Essays in Applied Econometrics

Dissertation submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by
Florian Schaffner
from Basel BS

approved in July 2017 at the request of
Prof. Dr. Rainer Winkelmann
Prof. Dr. Roberto A. Weber

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zürich, July 19^{th} 2017

Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgements

While I alone am responsible for this thesis, I would like to thank a number of people that have accompanied and supported me throughout the years at the Department of Economics at the University of Zurich. First and foremost, I would like to express my gratitude to Rainer Winkelmann, my thesis advisor. I have enormously benefitted from his input relating both to research and education, and I cannot thank him enough for his support and patience. My grateful thanks are extended to Roberto Weber for many critical comments which improved my work and for co-advising this thesis. I am indebted to Steven Stillman, who supported and guided me along the way.

With Johannes Kunz, Alessandro de Carli, Chloé Michel, Jean-Michel Benkert and Peter Hoeschler I had enthusiastic and critical fellows. Our discussions have led to fruitful collaborations and valuable friendships; for both I am grateful. I admire and value each and every one of them.

Finally, I thank my parents and my siblings for supporting me, encouraging me and giving me the freedom to pursue my academic aspirations. I thank Adriana, whom I value more than anyone for her infinite understanding and ongoing support.

Contents

Acknowledgements **iii**

List of Tables **x**

List of Figures **xii**

1 Introduction **1**

2 Swipe right - Preferences and outcomes in online mate search **11**

2.1 Introduction 12

2.2 The smartphone application 17

2.3 Measuring attractiveness and selectivity 20

2.4 Preferences 22

2.4.1 Estimation 24

2.4.2 Data on decisions 26

2.4.3 Results on preference parameters 28

2.5 Characterizing the initial match 32

2.5.1 Model 33

2.5.2 Empirical results on best ranks 40

2.5.3 Validating the model 43

2.6 Match progression 49

2.6.1 Opportunity sets 50

2.6.2 First impressions and final matches 53

2.7 Conclusion 56

References . . . 57
Tables and figures . . . 60

3 Information transmission in high dimensional choice problems: The value of online ratings in the restaurant market 79

3.1 Introduction . . . 80
3.2 Information transmission and the role of internet data . . . 83
3.3 Data . . . 86
3.4 Model . . . 89
3.4.1 Restaurant choice with perfect information . . . 89
3.4.2 Restaurant choice with imperfect information . . . 90
3.4.3 Parameters of interest . . . 94
3.4.4 Grouping restaurants by quality . . . 96
3.4.5 Estimation . . . 97
3.5 Main results . . . 99
3.6 Robustness checks . . . 102
3.7 Discussion . . . 104
References . . . 106
Tables and figures . . . 109

4 Predicting US bank failures with internet search volume data 119

4.1 Introduction . . . 120
4.2 Previous literature . . . 122
4.3 Data sources, data properties and descriptive statistics . . . 124
4.3.1 Data sources: Google Insights for Search . . . 125
4.3.2 Additional data sources . . . 127
4.3.3 Variables and summary statistics . . . 128
4.4 Model and results . . . 130
4.4.1 Model . . . 130
4.4.2 Results . . . 132

4.5 Conclusion 135
References 137
Tables and figures 141

A Appendix: Chapter 2 **153**

B Appendix: Chapter 3 **167**

C Appendix: Chapter 4 **169**

Curriculum Vitae **175**

List of Tables

2.1 Summary statistics on user-candidate attributes in decisions 60

2.2 Fixed effects logit results on preference estimates (coefficients) 61

2.3 Common vs individual preferences decomposition 62

2.4 Best achieved rank explained by search length, attractiveness and acceptance rate 63

2.5 Testing the uniformity assumption 64

2.6 Choices and matches 65

2.7 Results on the expansion of the opportunity set (all observations) 66

2.8 First impressions in later stages, females 67

3.1 Summary statistics 109

3.2 Cumulative checkins over time 110

3.3 Estimated parameters of the Dirichlet distribution 111

3.4 Cumulative checkins over time, aggregated by rating 112

3.5 Single period checkins over time 113

3.6 Cumulative checkins over time, higher grouping levels 114

4.1 Overview of samples used 141

4.2 Bank failures over time 142

4.3 Summary statistics 143

4.4 Results on survival and Google data availability 144

4.5 Main results, uncensored Google series only 145

4.6 Forecasting 146

A.1 Summary statistics on binary HI/BYE decisions, by gender 154
A.2 Linear probability model results on preference estimates 155
A.3 Fixed effects logit results on preference estimates (marginal effects) 156
A.4 Fixed effects logit results on preference estimates (robustness I) 157
A.5 Fixed effects logit results on preference estimates (robustness II) 158
A.6 Robustness results on best rank: all users 159
A.7 Robustness results on best rank: $N \geq 100$ 160
A.8 Results on median rank (all observations) 161
A.9 Search length descriptives 162
A.10 First impressions in later stages, males 163

B.1 Overdispersion in restaurant checkins at different levels of aggregation . . . 168

C.1 Description of Variables 170
C.2 Google Query Index Value Growth Rates 172
C.3 Additional results (censored Google series) 173

List of Figures

2.1 Subgame decisions leading to a match 68

2.2 Attractiveness $a = \frac{liked}{liked+disliked}$, by gender 69

2.3 Acceptance rate $s = \frac{likes}{likes+dislikes}$, by gender 70

2.4 Acceptance rates vs attractiveness, by gender 71

2.5 Outcomes as limit cases, by gender 72

2.6 Probabilities of different ranks across subperiods 73

2.7 Number of registered users, by month 74

2.8 s_r lower bound as measured by predicted ranks (conditioned on $Y = HI$) . 75

2.9 Probability of a match (attractiveness×acceptance rate), by gender 76

2.10 Distribution of inclusive values, by gender 77

2.11 Distribution of rank inclusive values by gender 78

3.1 Restaurant visits and ratings as of 2015 115

3.2 Dirichlet distributions for a set of different δ 116

3.3 Dirichlet parameters for different rating categories, over time 117

4.1 Google Search Volume Index and share price for "Washington Mutual" . . 147

4.2 Google search queries in the weeks prior to failure 148

4.3 Key balance sheet positions before failure, conditioned on observing uncensored Google series 149

4.4 Smoothed hazard rate estimate 150

A.1 Regional interest for BLINQ as measured by Google Trends data (1/2013 - 7/2015) 164
A.2 Age range of candidates in years, by gender (local polynomial fit) 165
A.3 Distance range of candidates in km, by gender (local polynomial fit) 166

Chapter 1

Introduction

This thesis contributes to a diverse set of current topics that may appear disparate at first glance, with chapters ranging from mate search behavior to restaurant ratings and bank failure prediction. Yet all chapters have a common denominator: They all make use of novel sources of online data, be it Google search volumes, ratings on internet platforms or decisions in a smartphone application. They are novel in that they originate from applications that did only emerge in the past decades. Typically, the datasets are large in that they cover millions of decisions at high granularity, making economic problems highly trackable in areas where data was previously scarce. Falling back on these data sources allows testing theoretical predictions, extending existing empirical findings in the respective literatures and shedding new light in corners that previously have been hard to assess empirically at all.

Chapters 2 and 3 are also related in that they both analyze individual choice problems in a context with many alternatives and limited information. Such setups and their analysis have existed for a long time, but have become more eminent with the expansion of the internet. In the pre-internet era, many choice sets were either limited or alternatives were only accessible at a significant search cost. With the advent of the internet, the number of alternatives has risen and the cost of acquiring information dropped, with the additional advantage from a researcher's perspective that behavior in these real-life examples can now by observed step by step. This thesis gives three examples; in this overview, I give a short summary of each.

The second chapter, titled *Swipe right: Preferences and outcomes in online mate search*, analyzes asymmetries in outcomes of online mate search. Matching women and men, the classic matching problem, has been hard to investigate up until recently, and the econometrics of matching still lag the developments in theory. This is largely due to restricted data availability: Most of the empirical research in matching analyzes datasets including only realized matches and cannot go beyond identifying match surpluses on the level of

a matched pair. Often, researchers have to work with limited or no information on how individuals end up in such equilibria. This changed with the rise of online-dating, which by 2010 has become the second-most important meeting channel for couples (the first channel being meeting through friends). It is likely to have become even more important since then: By 2012, the first mobile dating applications came to market, which in terms of active users quickly overtook desktop online dating.

The seminal contribution in the field of online dating is Hitsch et al. (2010), who analyze data from an online dating company to estimate preferences and assess the efficiency of online dating by assuming no search costs and comparing outcomes to theoretical predictions by the Gale-Shapley algorithm. By looking at data from the Swiss mobile dating application BLINQ, I build and extend on their contribution.

In a first step, I estimate preferences by analyzing binary willingness-to-date decisions to then construct individual rank-orderings of all the candidates an individual has (virtually) met. Here, I proceed similarly as in Hitsch et al. (2010), estimating a fixed effects logit model with individual-specific reservation values. Results highlight the importance of information captured in profile photos, a factor that often had to be ignored in previous research. Results also show a tendency towards homogamy, validating previous research and providing evidence that the frequently observed assortative mating patterns in marriage markets are at least in part due to preferences and not exclusively to search frictions.

Using the previously obtained rank ordering and ranking the most preferred candidate first, I then look at the lowest rank of all the candidates one has matched with and investigate asymmetries in these ranks across gender. Such asymmetries have not been studied previously or even been excluded by a common preference assumption, but they turn out to be quite significant: Females achieve a median rank of 8, while the same rank for males is 79.

In order to explain these asymmetries, I employ the theoretical framework of the Secretary Problem, an optimal stopping problem originating from the 1960's, extended to the

two-sided case in 2008. The model offers the advantage that individuals only have to rank partners they have actually met, rather than having an order over the whole universe of potential candidates. Achieved ranks are set into relationship to search length, own attractiveness and own selectivity, where search length is predicted to have a negative effect on the achieved rank, while own attractiveness and selectivity are expected to have positive effects.

The most striking finding in the analysis is that search length has no impact on females' achieved ranks. This roughly corresponds to the one-sided limit case of the problem, which is optimal from a female's perspective. Vice-versa, males' outcomes correspond to the pessimal case. Summary statistics on user behavior released by competitor applications such as Tinder or reported in previous research indicate similar asymmetries in other dating applications. With a general trend of couples marrying later and individuals searching longer, results suggest that these asymmetries across gender are likely to increase. This may turn out to be important, as other research suggests that imbalances within pairs increase the probability of a divorce further down the line.

As I show in the last part of the paper, the asymmetries are somewhat attenuated in later stages by males behaving more proactively when contacting females. The set of matched partners exhibit the same asymmetric patterns as after the initial match, and subsequent decisions to pursue a partner after an initial match are in line with first stage preferences. But as males tend to approach their more favorited matches, they manage to reduce asymmetries to some extent. The findings might also translate to similarly structured markets, most notably job search, where recruiters (candidates) face a large set of candidates (job postings) but can only search through a subset of potential candidates (jobs) before making a hiring decision.

Chapter 3 is titled *Information transmission in high dimensional choice problems: The value of online ratings in the restaurant market* and looks at the econometric estimation

of choices in the presence of social interactions. The data is sourced from the online urban guide Yelp, providing visit numbers and ratings of restaurants.

From the perspective of an individual, choosing a restaurant is a demanding problem in two dimensions: On the one hand there is an abundance of alternatives. On the other hand, information about the quality of restaurants is limited prior to one's choice. One can think of two ways to address these issues. One is to provide more information. There has been much debate about whether ratings of the type provided by Yelp represents a new kind of social currency and a solution to the informational problem, and I contribute to that discussion in trying to assess the informational value of these ratings. The second remedy goes in another direction, even arguing that more information might worsen the problem in the context of high-dimensional choice sets (Simon, 1955). Instead, it is argued that individuals resort to social interactions and orient themselves by the choices of others, either because of a direct effect that the company of others increases one's utility or because the choices serve as a signal of quality. Either way, such interactions can lead to the highly skewed outcome distributions seen in the restaurant market, and have been shown to impact outcomes in various contexts such as book sales, movie attendance or the choice of health care plans (Beck, 2007; De Vany and Walls, 1996; Sorensen, 2006).

Taking social interactions into account econometrically is challenging, especially in the context of cross-sectional data on the restaurant level (rather than a panel structure with individual-level data). What I do to embed the social dynamic in the choice problem is to resort to a Dirichlet-multinomial distribution. Contrary to the classic multinomial logit model, the Dirichlet-multinomial treats the probability vector in the multinomial as a random variable rather than as a constant, allowing the probability of choosing a restaurant to rise in proportion to the number of previous visitors and enabling a rich-get-richer-dynamic. Conveniently, the multinomial logit is nested in the Dirchlet-multinomial model, which makes testing one model against the other readily implementable.

The restaurant outcomes are measured by checkins and the corresponding probability

vector is modelled as a function of a restaurant's rating, while the dataset is segmented into more than 200 markets defined on a ZIP code and price category level. The strength of social interactions is modelled as a function of group-level variables including the number of competitors, the number of reviews and a proxy for income inequality.

Results show that correlation across choices of restaurant guests is present and can be modelled as a function of group-level variables related to information exchange. Ratings and other factors such as price categories do play a role, though traditional modelling approaches that ignore social interactions tend to overstate its importance both in terms of statistical and economic significance. The presence of correlation across choices mainly means that predictions of outcomes become much more uncertain. Being mostly a cautionary note about inferring too much from ratings, the conclusion may not be particularly attractive; it is still a noteworthy one.

Chapter 4, the final chapter, is titled *Predicting US bank failures with internet search volume data* and investigates how well weekly Google search volumes track and predict over 400 bank failures in the United States between 2007 and 2012. While there exists an extensive literature prediction bank failures via balance sheet and revenue positions, the timing of failures itself is hard to predict in the absence of share prices.

The approach taken in the chapter is to use Google search volume indices as a high-frequency proxies for public attention. The indices are a reflection of the number of Google search queries for a particular term on a weekly basis; anecdotal evidence indicates that contrary to for example news headlines, Google search volumes closely track share prices of banks listed on an exchange and can therefore serve as a substitute. Google data has been used in different contexts to predict or nowcast events of both economic and non-economic nature. Economic examples include unemployment rates, consumer behavior, housing prices or inflation (Askitas and Zimmermann, 2009; D'Amuri and Marcucci, 2010; Choi, 2009; Tefft, 2011; Guzman, 2011; Choi and Varian, 2012; Goel et al., 2010; McLaren

and Shanbhogue, 2011; Wu and Brynjolfsson, 2013). Non-economic examples cover epidemics, kidney stone incidences or suicide risks (Ginsberg et al., 2008; Breyer et al., 2011; McCarthy, 2010).

I use a sample of both failing and non-failing US banks between 2007 and 2012 and complement the Google indices with FDIC-provided data on balance sheet and revenue positions, where the selection of these positions is guided by previous studies. To predict the failures, I estimate different duration models with time-varying covariates and piecewise-constant hazard rates. Google indices vary on a weekly basis, while balance sheet and revenue variables vary by quarter.

Higher Google search volumes go hand in hand with higher failure rates, and the estimated coefficients for the Google volume growth index are highly significant. Effects on the hazard rate effects are comparable to positions labelled as troubled assets in the balance sheet. However, Google volumes typically rise only two or three weeks before bank's failure, with Google's predictive power quickly dissipating for failure rates further in the future. Balance sheet and revenue positions, on the other hand, typically follow longer-term trajectories. Incorporating search volumes in bank failure predictions should therefore primarily be seen as en extension to existing models, not a replacement thereof.

In summary, all chapters make use of data sources that are likely to become more prevalent in future research. All of them track or proxy processes and phenomena that were otherwise hard to pin down at such detailed levels, be it search behavior and matching, word of mouth and social interaction dynamics or public interest in a specific company. They may serve as add-ons to existing econometric models as well as substitutes to them. And while in many cases, they only cover subsamples of a population, they have nevertheless become large and popular enough to be interesting ecosystems to analyze within themselves as well as to have real-life impacts: Online dating has risen to become the second most important meeting channel for couples; the number of monthly Yelp users is more than 15 times larger than the Swiss population; and Google has become everybody's go-to-gateway

to access information.

Chapter 2

Swipe right - Preferences and outcomes in online mate search

Acknowledgements: I would like to thank Jan Berchtold, Rema Hanna, Johannes Kunz, Edward Lazear, Janina Nemitz, Andrew Oswald, Steven Stillman, Roberto Weber, Rainer Winkelmann, Michael Wolf, Alex Zimmermann, participants at the Zurich Workshop on Economics 2016, participants at the 2017 Grindelwald Personnel Economics Workshop and seminar participants at the University of Zurich for helpful comments and suggestions. I am indebted to Alessandro De Carli, who assisted me throughout this project. Errors and omissions are my own.

2.1 Introduction

Over the past 20 years, online mate search has become one of individuals' main intermediaries to find a partner, and previous research has successfully linked online dating patterns to observed patterns in the overall marriage market (Hitsch et al., 2010). Inexistent up until the mid 1990's, more than 20 percent of heterosexual couples met online in 2010; for same-sex couples, that fraction was almost 70 percent (Rosenfeld and Thomas, 2012). Shares are likely to have increased since then, with the use of online dating or mobile apps by young adults nearly tripling between 2013 and 2015 (Smith, 2016) and mobile dating overtaking traditional online dating in 2012 (Sales, 2015). Meanwhile, other channels for mate search such as meeting through coworkers, family or at college are in a steady decline. This shift opens interesting research opportunities, because as opposed to other mate search intermediaries, the online channel offers a significant advantage: trackability.

Starting with Becker (1973), preferences and resulting outcome patterns in mate search and the matching literature more generally have gotten a lot of attention in both theoretical as well as in empirical research. But empirical work has lagged the theory, mainly because applied research in mate search and matching more generally had to be content with datasets only containing realized matches, with no information regarding how individuals ended up in an equilibrium (e.g., Lee, 2015). More recently, this restriction on the data side lead to advances in the econometrics of matching models. Still, empirical work on search and matching is tightly embedded in theory and relies on strong assumptions as long as only sets of realized matches are available (for a survey, see Chiappori and Salanie, 2016).

This paper works with a rich online dating dataset tracking mate search behavior from start to finish: The potential partners an individual considered, realized as well as rejected matches, the choice set of available partners, and final outcomes. The setup allows to circumvent many of the potential econometric challenges, answering questions such as:

What is the ideal partner of an individual? How successful are individuals in matching with that ideal partner? To what extent does that success depend on their search intensity, their own behavior as well as the behavior of candidates on the other side? After an initial match, are actions at a later stage consistent with the early-stage decisions of an individual?

The first contribution of this paper is its focus on asymmetries in behavior and outcomes across gender in the mate search problem. Taking the individual-specific rather than the pair-specific view allows conclusions about the optimality of an equilibrium matching from a new perspective. Such considerations might be important: With a large number of participants in a matching market, the number of stable matchings increases considerably (Pittel, 1989). But not all of these matchings are equally preferable from an individual viewpoint, which may have consequences for longterm prospects of a couple. It has been shown that asymmetries within pairs result in higher divorce rates (Guven et al., 2012), while trends such as marrying later may increase asymmetries across females and males (Kashyap et al., 2015). Previous research either had to ignore asymmetries, because data structures only allowed identification at the level of a matched pair, or deemed them unlikely by assuming common preferences. To the extent that they were taken into account, asymmetries were attributed to exogenous factors such as uneven sex ratios, whereas I will also allow asymmetries to emerge endogenously.

The second contribution of this paper is the estimation of preferences. Knowing mating preferences helps understand the causes of the observed assortative patterns in marriage, which in turn affect economic variables of interest such as income inequality (Burtless, 1999). In a seminal contribution, Hitsch et al. (2010) analysed preferences using data from an online dating platform. In their setting, users browse online profiles of potential mates, and, if they find the information provided on the profile appealing, send out a first-contact e-mail. The authors use the binary decision "Email yes/no" in order to estimate a model of revealed mate preferences. They find, for instance, that both men and women have a strong preference for similarity and, famously, that women have a stronger preference for

income relative to physical attributes than men.

This paper takes the Hitsch et al. (2010) approach to the next generation of dating technology, namely mobile apps. A key difference between computer-based online dating and mobile app-based dating platforms such as Tinder is that users cannot freely browse through profiles, but need to respond to externally selected proposals. In contrast, preference estimates in Hitsch et al. (2010) only use decisions from a pre-selected set of choices, where the selection is made by the user based on his or her preferences, and thus endogenous. This pre-selection issue is avoided in the application studied here — effectively attributing any emerging assortative patterns to preferences rather than endogenous meeting opportunities. Also, as opposed to starting or replying to a conversation, individuals are forced to take independent decisions without any interaction with the candidate, thereby excluding any potential endogeneity issues by design. Transaction cost are even lower for mobile apps than they already are for traditional dating platforms, and strategic behavior, found unimportant by Hitsch et al. (2010), are even less of a concern. Finally, individuals put a lot of weight on candidates' photos in their decisions[1], a factor that could only be partially accounted for in Hitsch et al. (2010), as only about a quarter of users posted at least one photo at all. On the downside, the information on potential dating partners by mobile dating apps is less rich. For instance, Hitsch et al. (2010) estimate the effect of income, height, body mass index and religious denomination on mate choice, whereas the decisions in the mobile application are mainly based on profile pictures and age.

The analysis of this paper follows the three steps of the dating process as it presents itself to the typical user of a dating app. Data originate from a Swiss location-based mobile dating app encompassing over 17,000 individuals making a total of more than 33 million decisions. In a first step, users have to make independent, binary willingness-to-date decisions for a random sequence of proposals. They can choose how many proposals to consider, and there is no limit on the number of acceptances. However, there is no

[1]The New York Times, Tinder, the Fast-Growing Dating App, Taps an Age-Old Truth, `http://nyti.ms/29WqO2e`

backtracking: if a proposal is rejected, it is not possible to change one's mind later. From these binary willingness-to-date decisions I estimate individual preferences over attributes of potential partners and rank all partners considered in an individual's "choice set".

Step two determines whether or not there is a match and how "favorable" such a match is. A match is defined simply as mutual acceptance ("Hi" responses) of the proposed mate, in which case a chat window opens. Based on my preference results, I can determine how close the actually matched partners are to the ideal partner (in terms of ranks), and how this distance depends on how attractive an individual is, how selective she behaves and how long she is willing to search for a mate. In particular, I show that such asymmetries in behavior and outcomes between men and women are closely related, taking a theoretical model of two-sided matching as guidance (the "Secretary Problem", Ferguson, 1989; Eriksson et al., 2008).

In a final third step, I analyze opportunity sets and follow chat messages to determine whether a telephone message is exchanged (which happens in 2 percent of initial matches), corresponding to a match as defined by Hitsch et al. (2010). Unlike in previous steps, in step 3 matched individuals are free to interact with each other as much as they want, introducing endogenous decision-making. Whereas in step one, individuals took snap decisions in just a few seconds, interactions in step three last longer and allow both mates to gain additional information about their matched candidate. By introducing the preference ordering from step one as a predictor for a phone number exchange, I can connect the short-term, snap judgment stage with the longer-term, endogenous interactions in stage three. Controlling for exchanged messages, I can test whether first-stage decisions are in line with later-stage interactions. As decisions in step one and step three are different events, predicting later stage matches serve as out-of-sample predictions of the estimated preference parameters revealed in step one.

Results on revealed preferences show that females as a group behave very selectively, whereas male acceptance rates are much more heterogeneous. Overall, physical attrac-

tiveness of a candidate is the primary factor in the willingness-to-date decision for both men and women, a result in line with previous research on online dating services. The age of a candidate is an additional important factor, with females preferring older and males preferring younger candidates. At the same time, both genders dislike age differences, counteracting the former age effects and resulting in a total effect with an inverse U-shape. Results generally show a strong preference for homogamy, with females and males disliking both positive and negative differences between a candidate and themselves. These tendencies confirm the assortative mating patterns that are well documented in previous research.

When looking at the ranks of the best-matched partner of each candidate, I find that females are getting on average more highly ranked partners than men. With median ranks of females at rank 8 and median ranks of males at 79, outcomes differ by a factor of almost ten. When analyzed in the model context, these outcomes suggest an almost female-optimal and male-pessimal matching. A female's achieved match rank does not depend on search effort, approaching to the one-sided limit case of the model. For men, more intensive search pays out in terms of rank percentiles, with absolute ranks growing at a slower pace than search length. I make the case that males approach another limit case of the model in which their payoffs converge to their respective outside options. Combined with the generally observed increase in the age at which individuals get married, this finding suggests that asymmetries within couples are likely to increase as both males and females search their partners over longer periods of time.

The ranks assigned in the first stage are in line with the phone numbers exchanged later, with lower ranks increasing the probability of starting a conversation, replying to a first message and exchanging phone numbers. This is reassuring, as it suggests that high-frequency snap judgments based on limited information in the first stage are consistent with actions taken at later stages. In the majority of cases, males make the first contact and aim high, contacting better-ranked females while ignoring their own ranking in the

female's eyes. Females show more reciprocal patterns in both starting a conversation as well as replying to a first contact, taking into account their ranking in the candidate's eyes.

This paper is structured as follows: Section 2.2 introduces the smartphone application. Section 2.3 presents preliminary statistics on key variables, while section 2.4 presents results on preferences. Section 2.5 introduces a theoretical framework for the first stage decision and analyzes empirical results in the context of that framework. Section 2.6 discusses opportunity sets and later-stage decisions. Section 2.7 concludes.

2.2 The smartphone application

Data is sourced from BLINQ[2], a Swiss location-based mobile dating app which first went online in 2013. The goal of the app is to match two persons, allowing them to chat and eventually meet for a date. Both the app and the app's competitors (e.g., Tinder) have become very popular in the dating life of young Swiss individuals. As measured by Google Trends data (Figure A.1 in the appendix), BLINQ is most popular in the German-speaking part of Switzerland, particularly in the canton of Zurich and its adjacent regions. The application and registration are free of charge.

Unlike traditional online dating websites (see, e.g., Hitsch et al., 2010), users in the BLINQ app are not free to browse through profiles. The only filters they can set on their choice set are filters on sexual orientation, age range and geographic distance. When a user opens the application on her smartphone, she is presented with an exogenously selected candidate from the set of candidates satisfying their search filter. Users cannot skip candidates, but are forced to take a decision on each candidate in order to move on to the next candidate.

As meeting opportunities are then externally assigned, this largely gets rid of the problem of disentangling dating preferences from endogenous meeting opportunities (i.e., search

[2] `http://www.blinq.ch`

frictions, Belot and Francesconi, 2013), which is of particular interest when studying assortative mating patterns. However, the ordering of presented candidates in the sequence is not fully random. The app's algorithm orders potential candidates by a combination of the response of the candidate (candidates who already positively responded to the user appear sooner to ensure timely notification of a match), activity (more active users appear sooner in the sequence), attractiveness (measured by the fraction of *HI*'s a candidate gets), the (standardized) difference in attractiveness between user and candidate, and the distance between user and candidate. This ordering process is executed each time a user opens the app and sends a query to the app's server. Preset filters may be overridden if the set of candidates fullfilling the restrictions is too small.

The ordering of candidates will be crucial in the theoretical model used in this paper — in particular, I will assume that the next candidate's rank is uniformly distributed, an assumption I will explicitly validate in a later section. I do not rely on meeting opportunities to be fully exogenous, however, as I abstain from drawing definitive conclusions about the drivers of assortative mating. The model itself assumes users only rank candidates they have actually been shown (as opposed to having a rank ordering relative to the whole candidate pool), but makes no assumptions about potential selection effects with respect to choice sets.

The user is given some information about the candidate, including first name, photos, geographic distance, school and mutual friends (see Figure 2.1(a)). Based on this information, the user has to decide whether to say *HI* or *BYE* to the candidate (swiping right or left on the phone's screen). I will call each of these *HI/BYE*-decisions a subgame or a period, using the terms interchangeably. I will also refer to the decisions as ratings, since the decision to date someone is also an indication of the attractiveness of the candidate.

Profile information is imported from Facebook to ensure credible information, and newly registered users are examined by the app's developers in order to avoid and filter out fake profiles. I only analyse data of users who passed and completed the registration process,

are at most 40 years old and were located in Switzerland at the time the data was drawn. The dataset was drawn in July 2015. I will focus exclusively on heterosexual mate search.

If a user says *BYE*, she moves on to the next candidate. The same thing happens if the candidate on the other side rejects the user. There is no backtracking: Once a user has rejected a candidate or has been rejected by a candidate, she can never revoke that decision. If both she and the candidate say *HI*, both users get notified about the match (see Figure 2.1(b)) and a chat window opens that allows them to exchange messages (see Figure 2.1(c)). Going forward, I will refer to the step in Figure 2.1(a) as the first step or early stage, the screen in Figure 2.1(b) as step two and the screen in Figure 2.1(c) as the third step or late stage. Throughout the paper, but with the exception of the last section, I will define a match as both user and candidate giving a positive response. I use alternative definitions in the section on later matching stages.

It is important to stress that at the time of the decision, the user has no information regarding if or how the candidate has decided on herself, which significantly simplifies the estimation of preferences. If both candidates are still interested after exchanging a few messages, they will usually exchange phone numbers through the chat, then exit the app and possibly meet in person.

On the downside, anything that happens beyond messaging in the application's chat is not registered. Also, matches in the application are not exclusive. A user can collect multiple matched candidates, which in turn may have multiple matches themselves. A match as defined by the app and in this paper should not be seen as an equivalent to being in a relationship (let alone marriage), but rather as a mutual signal of interest and an opportunity to go on a date with someone. As such, the application refers to the earliest stage of a relationship and is more akin to speed dating.

2.3 Measuring attractiveness and selectivity

The dataset comprises observations on 6,066 females and 11,302 males, resulting in a gender ratio of almost 2:1. That ratio stays roughly constant over time. I dropped any users that have not completed the login process, have not passed the application's screening process or have been blacklisted, thereby filtering out fake profiles. I also dropped the homosexual and bisexual users in the data due to their limited number. I do keep bisexual individuals as candidates when evaluating preferences, meaning that heterosexual users can rate bisexual candidates but not vice-versa.

With respect to search effort, females take 1,695 decisions on average, compared to 2,032 for males.[3] I will use the number of decisions as the measure for search length. More than 99.9 percent of females and 92.5 percent of males do not rate the whole set of candidates, making the constraint of a finite candidate pool not binding for a large majority of users. In other words, although the 2:1 gender ratio mentioned above might sound extreme, the dating pool on the app is deep enough to make that ratio effectively irrelevant for the majority of users.

I define the overall attractiveness of a candidate as the fraction of positive responses (HI or $likes$) a user gets — in other words, the average probability of a user to be accepted by a candidate. In later estimations, this overall measure captures any characteristics that are not included as a covariate (e.g., age). Given the application's setup, it is reasonable to assume that it mostly captures information transmitted through photos, where the information in the photo may be directly related to a candidate's physical appearance as well as surroundings.

The assumption is backed up by previous research: Working with the same data, Rothe et al. (2015) extracted visual features from candidate's photos to predict willingness-to-date decisions. Out-of-sample predictions based on one photo alone were shown to be correct in

[3]Median values: 1,103 for females, 1,213 for males.

more than 75 percent of cases, with improved accuracy if a user's decision history was taken into account.[4] The authors' algorithm was further validated using the dataset in Gray et al. (2010) where subjects were asked to judge facial beauty of candidates in photos. Overall, these findings suggest that photos play a major role in all individuals' decisions, which in turn will be reflected in the attractiveness measure.

There are strong differences in attractiveness measures between females and males. Females have an average attractiveness of 0.486, indicating that roughly every second time a female is shown to a male user, she gets a positive response. The overall distribution of females follows a bell-shaped beta density shown in Figure 2.2, resembling previously found patterns (Rudder, 2014). Its unimodal shape itself has been highlighted in previous research, as one could imagine for example bimodal, beauty-and-the-beast like distributions as well.

Females rate males' attractiveness much more conservatively, with the average male attractiveness at 0.072 or approximately 7 percent, and the overall distribution in Figure 2.2 skewed to the right. Again, previous research shows similar patterns (Rudder, 2014). Both attractiveness measures are close to the same measures on the US application Tinder, which is reported to be 14 percent (females) and 46 percent (males), respectively.[5] If I look at the candidate's average attractiveness measure on a decision level rather than on the level of an individual, these numbers are even closer (females: 14 percent; males: 50 percent), suggesting that individuals behave similarly across these applications.

I define the acceptance rate or selectivity measure of a user as the number of times a user gives a positive response, divided by the total number of responses. In the aggregate, this roughly corresponds to the attractiveness measure of the opposite gender, though not exactly, as users who stay on the application longer will also be shown more often.[6]

[4] A demonstration of their algorithm can be found at `http://www.howhot.io`, where visitors are free to upload their own photos and get an estimate of the facial attractiveness of the person depicted.

[5] Source: The New York Times, Tinder, the Fast-Growing Dating App, Taps an Age-Old Truth, `http://nyti.ms/29WqO2e`

[6] If every user rated every candidate and vice-versa, the attractiveness measure and the acceptance rate would be equal.

The mean value is 0.116 for females and 0.506 for males. Aside from the previously cited statistics, females being choosier than males is observed in other dating contexts as well (e.g., Belot and Francesconi, 2013; Fisman et al., 2006).

An interesting pattern can be seen when looking at the whole distribution in Figure 2.3: Whereas the distribution of the acceptance rate of females roughly resembles the distribution of male attractiveness and shows homogeneous, relatively concentrated behavior within the group of females, male selectivity is much more evenly distributed, resembling a uniform distribution rather than the bellshaped distribution of female attractiveness.[7] In particular, there exist some very selective male users as well as a group of males that is willing to accept almost any candidate. As discussed later, the model employed in this paper offers an explanation for how these differing behavior patterns may arise. Both distributions remain largely unaffected when conditioned on the users' attractiveness measures.

On the individual level, there is a strong interplay between a user's attractiveness and her own acceptance rate. The relationship is shown in Figure 2.4.[8] The more attractive a user, the more selective she behaves (note that more selective behavior means a lower acceptance rate). Such behavior makes sense if users target a finite, manageable number of matches rather than maximize total matches (see also Table 2.6). The negative relationship between own attractiveness and own acceptance rate is one of the fundamental theorems derived in the theoretical model employed in this paper.

2.4 Preferences

To be able to say anything about whether mate search is successful, I need to know what individuals consider attractive. To what extent is the ideal partner a universal type? Do individuals only care about the attributes of a mate itself, or is the difference in these

[7] One may also argue that the male acceptance rate density is a bimodal distribution with the second mode at the boundary of 1, hinting at a mixture between two types with different levels of acceptance rates.

[8] The graph for males excludes one outlier, the most attractive male (attractiveness=0.75). Including the observation leads to a stronger uptick at the right end of the attractiveness scale.

attributes with respect to oneself relevant, too? This section estimates preferences for males and females, revealed by binary HI/BYE-decisions (yes/no) made in the first stage in the application. Revealing preferences serves three purposes: For one, by knowing preferences, I can construct individual rank orderings that allow me to analyze outcomes. With the ideal candidate of an individual ranked first, looking at the ranks of final matches gives an indication of how close an individual's matched mate gets to her ideal partner.

Second, several aspects of preferences are interesting in their own right. For one, it is a priori unclear how common or idiosyncratic preferences are across individuals. In biology, researchers typically assume common preferences, with agents evaluating their preference for a mate according to a universal measure such as physical fitness. In such a case, rank orderings are identical across individuals, leading to a unique stable equilibrium. At the other end of the spectrum are independent preferences, where the preference ordering of one individual is fully independent of the ordering of another individual. By introducing both common elements as well as pair-specific variables and fixed effects, I can draw some conclusions on the relative importance of common and individual preference components.

Third, I also shed some light on the discussion of assortative mating, i.e., the frequently observed pattern that individuals mate with partners that resemble them across different dimensions (e.g., young with young, high income with high income, high education with high education). Identifying the drivers behind assortative mating has been a challenging task for empirical researchers, as common preferences, a preference for homogamy as well as endogenous meeting opportunities (search frictions) offer explanations for the observed pattern. Given that I observe a large and exogenously imposed choice set in my data and search costs are minimal, I can reasonably assume away endogenous meeting opportunities. In other words, any observed assortative patterns are likely the result of common and individual preferences. Note that this is not to say that endogenous meeting opportunities and other search frictions might not enforcen such patterns.

2.4.1 Estimation

Preferences are revealed through the introduction of latent random utility functions, where I use the assumption that if a user is willing to match with candidate j but not with candidate k, she prefers a potential match with j over a match with k. Utilities are assumed non-transferable. Let $U_W(m, w)$ be the expected utility that a female user w gets from a potential match with male user m, and let $\nu_W(w)$ be the reservation utility w gets from her staying single and continuing the search for a partner (in other words, the outside option). She chooses to say HI to a candidate in period r if and only if

$$U_W(m, w) \geq c(w, r).$$

The cutoff value $c(w, r)$ is both individual specific and time dependent[9]. As in the limited-awareness model in Menzel (2015), the utility of a match with a candidate the user never meets is set to minus infinity. I will use the female perspective for the remainder of this section; utilities for males are defined analogously.

Given this threshold-crossing decision rule, mate preferences can be estimated using standard discrete choice models. A woman's utility is defined as a combination of deterministic, observed attributes of candidate m as well as woman w, a parameter vector θ_W as well as an idiosyncratic term, $U_W(m, w) = U_W(X_m, X_w; \theta_W) + \varepsilon_{wm}$. As in Hitsch et al. (2010), I split the attribute vector and parameter vector into separate components: $X_m = (x_m, d_m), \theta_W = (\beta_W, \gamma_W^+, \gamma_W^-, \vartheta_W)$. The latent utility of woman w from a match with man m is parametrized as

$$\begin{aligned} U_W(X_m, X_w; \theta_W) =& x_m' \beta_W + (|x_m - x_w|_+')^\alpha \gamma_W^+ + (|x_m - x_w|_-')^\alpha \gamma_W^- \\ &+ \sum_{k,l=1}^{N} \mathbb{1}\{d_{wk} = 1 \text{ and } d_{ml} = 1\} \cdot \vartheta_W^{kl} + \varepsilon_{wm} \end{aligned} \tag{2.1}$$

The first component in the above equation captures common preferences for a male can-

[9]Potential time dependencies are discussed in more detail in the model section.

didate's attributes, regardless of a woman's own attributes. By taking differences between the attributes of a candidate and a user, the second component captures pair-specific (i.e. individual) preferences. Negative parameters indicate a preference for assortative mating, i.e. users prefer candidates resembling themselves over candidates that differ from them. I estimate the parameters for positive and negative differences separately, allowing for positive differences to have a distinct impact from negative differences by splitting them up into two separate parts with parameter vectors γ_w^+ and γ_w^-, respectively. In order to circumvent identification issues, the differences are exponentiated to the power α (throughout this paper, $\alpha = 2$). The summand collects indicators equal to one whenever both user and candidate share an attribute (e.g., both speak German, both have a university degree), capturing additional pair-specific characteristics. The third component embeds a user-specific fixed effect for a user's own characteristics as well as an idiosyncratic term. Finally, I control for the effect of time in $c(w, r)$ by including a period variable r in the estimation.

To estimate the model, I assume that ε has the standard logistic distribution and is i.i.d. across all pairs of men and women and estimate a individual fixed effects logit model. Reservation values $\nu_W(w)$ and $\nu_M(m)$ are estimated as fixed effects. Note that both reservation values and a user's own attributes are captured in the fixed effect. Choice probabilities are defined as

$$\Pr(w \text{ gives } HI \text{ to } m) = \frac{\exp(U_W(X_m, X_w; \theta_W) - c_{wr})}{1 + \exp(U_W(X_m, X_w; \theta_W) - c_{wr})}. \tag{2.2}$$

It should be pointed out that independence across partners and from observed characteristics is a strong, but standard assumption in the matching literature that makes estimation of the model straightforward (Chiappori and Salanie, 2016). Note, however, that in the present setup, decisions between two matched candidates are indeed independent, as both user and candidate learn about the other's decision only after they made theirs and are not allowed to interact until they mutually agree to get matched.

2.4.2 Data on decisions

For computational reasons and to avoid giving to much weight to heavy users, the estimation of preferences only uses the first 100 decisions of every user in the application. These 100 decisions cover roughly 8 percent of a user's search length, on average. Summary statistics on decisions and the users involved are provided in Table 3.1. Note that the variables mostly relate to the users' attributes, not the candidates' characteristics (apart from differenced measures). Standardized measures are normalized by gender and over individuals (as opposed to decisions) — as some individuals show up more often in decisions than others, reported summary statistics may deviate from the expected mean of 0 and variance of 1.

Variables can be grouped into three segments: physical attractiveness, demographic and socioeconomic characteristics, and a geographic variable. Physical attractiveness is measured by the *attractiveness* variable[10], which is defined as previously. Note that a user's own decision has been calculated out of candidate's attractiveness measures in order to avoid endogeneity issues. Also note that the *attractiveness* measure is not observed by users, but only by the researcher.

In order to measure differences in attractiveness in a pair, the measure has been standardized within gender. Summary statistics in Table 3.1 indicate that the sampled decisions include slightly above-average candidates with respect to attractiveness. The table also lists the *acceptance rate* mentioned previously, a behavioral variable not included in the regressions but captured in the fixed effect.

Age is high up in the list of important demographic and socioeconomic variables. *Age* is measured in years and balanced between males and females. All users are between 13 and 40 years old, covering the prime age range for dating. Age 13 is the minimum age to register on Facebook; I dropped the few people above 40 years old. When estimating preferences, I will introduce a cutoff at age 18 and measure the absolute distance in years

[10]Differences in HI and attractiveness measures are due to the capped sample after the first 100 decisions as well as dropping incomplete, hidden and blacklisted profiles.

from 18.[11]

University is a dummy indicating whether the user lists a university in the education section of her Facebook profile, whereas *Both university* indicates whether both user and candidate have listed a university on their profiles. Males report slightly higher university rates than females, but differences are not statistically significant. *Same school* indicates whether user and candidate have been at the same school.

German speaking indicates that the applications language is set to German, with the assumption that the language setting is an approximate indicator for the main language spoken by the user. *Both German speaking* is equal to one whenever $German = 1$ for both user and candidate.

On a more social dimension, *no.of friends* is the number of Facebook friends, measured in hundreds. *Mutual friends* is the number of mutual Facebook friends that also use the dating app, while the squared differences in friends is the difference in Facebook friends, measured in units of 100,000. All of these measures are included to capture the sociability of a person, with more outgoing individuals having a higher friendcount, which supposedly has an effect on how likely a candidate is to accept and contact a user. Note that only the mutual friends variable can be directly seen by the user, an indication of how much the pair's social circles overlap.

Distance is the distance in km between user and candidate, calculated using longitude and latitude coordinates. Note that these coordinates were only drawn once when the dataset was compiled, as the app does not record geolocational data for every single decision of a user. Implicitly then, I assume that users do not move. Decisions for users and candidates that were more than $300km$ apart were dropped to avoid including potentially fake user profiles while at the same time ensuring that users roughly within the borders of Switzerland stay in the dataset.

Finally, TRX records the decision number in the sequence, capped at 100 (in other

[11]The application itself sets age filters that separate minors from adults, which is why I introduce this cutoff.

words, it is the censored measure of the number of decisions in Table 2.6). The model (discussed later) predicts that users should get less selective as they approach the end of their search, which is why it is important to take the time factor into account when estimating preferences.[12] The TRX variable averages below 100 as some users quit before taking 100 decisions.

2.4.3 Results on preference parameters

Preferences for females and males are estimated separately. Table 2.2 and Table A.3 in the appendix present results on the fixed-effects logit model, with Table 2.2 listing coefficients and Table A.3 listing marginal effects.[13] As a robustness check, I estimate the model with 100 randomly drawn HI/BYE-decisions of all users in case the first 100 decisions lead to different results than 100 randomly drawn decisions of a user. I also run a robustness check by drawing the full search history of a limited set of randomly drawns users. The results on both robustness checks are presented in Tables A.4 and A.5, complemented by Table A.2 presenting results of a linear probability model as a baseline specification. Results are largely equivalent to the results shown here.

I also look at potential strategic behavior, i.e. whether an individual does not give a positive response because she anticipates the candidate would decline. This could potentially confound estimated preference parameters, as discussed in more detail in Hitsch et al. (2010). I proceed as in Hitsch et al. (2010), including a covariate inversely proportional to the candidate's acceptance rate, $pr = 1/accrate$, in estimation. Note that the candidate's acceptance rate is not directly observed by the user herself. Results show that although the coefficient on that strategic variable is statistically significant, including or omitting it does not alter the remaining preference parameter estimates in any meaningful way.

[12]Note that while the TRX variable is included in the estimation of preferences, it is ignored when calculating rank orderings as these rankings are not time-dependent.

[13]The calculation of marginal effects relies on the assumption of a fixed effect of zero and therefore should be interpreted accordingly.

In all tables, the first two columns present estimates on female preferences, whereas columns 3 and 4 present estimates on male preferences. The signs of the effects can be directly interpreted from the coefficients presented in Table 2.2. With respect to variables that measure differences, a negative coefficient can be interpreted as a preference for likeness or similarities, whereas positive coefficients indicate a preference for dissimilarities.

According to the results of the linear probability model in Table A.2 as well as marginal effects listed in Table A.3, attractiveness (as well as the differences within a pair) is the most important factor affecting the probability of saying HI to a candidate. As expected, the attractiveness measure has a strong positive impact on the likelihood of a positive rating on a candidate for both genders. Perhaps more surprisingly, differences in attractiveness are generally disliked in either direction, with marginal effects calculations hinting at stronger effects in the case where the user is less attractive than the candidate. On the individual level, the positive coefficient on the attractiveness of a candidate in combination with negative coefficients on differences leads to a u-shaped total effect of attractiveness, with its peak at the user's own attractiveness level.

This suggests a decisive role for physical attractiveness and other visual cues reflected in the profile photos. Contextualized and following intuition, this result is perhaps not surprising, as the application as well as its competitor apps are built around photos, and online dating companies themselves have become increasingly aware that looks are the most signficant factor in willingness-to-date-decisions, whereas other features such as common interests or education only play a secondary role. In the words of the online dating platform OKCupid, "a person's profile picture is worth that fabled thousand words, but your actual words are worth... almost nothing."[14] In light of the fact that the median time to take a decision in the app is approximately 5 (females) and 3 seconds (males), the suggestion that individuals decide mostly on the basis of photos is not only plausible, but supported by previous psychological research on first impressions (Willis and Todorov, 2006). At

[14]Source: The New York Times, Tinder, the Fast-Growing Dating App, Taps an Age-Old Truth, `http://nyti.ms/29WqO2e`

the same time it is an important finding, as other studies ignored visual features in their estimations. In the paper by Hitsch et al. (2010), for example, only 27.5 percent of users post a photo at all.

Age is another significant factor in the users' willingness-to-date decisions, and what appears supported by anecdotal evidence is confirmed: Males prefer younger partners, whereas for females it's the opposite.[15] Specifically, males prefer females about 3.5 years younger than themselves, while females prefer males that are 1.8 years older. Taken together, the respective preferences should lead to couples where men are older than women. Again, though, the effect is U-shaped: Slight age differences are preferred, but as the age gap with respect to an indivduals own age widens, there is a point at which the female (male) preference for an older (younger) is overturned by a preference for a similarly aged partner.

Other factors included in the estimation only move the needle compared to the effects of attractiveness and age, if they are significant at all. The two estimated effects on university education are positive with the exception of the university dummy for a female candidate presented to a male, though none of these effects are statistically significant at the 5 percent level. The effect of being at the same school is positive but only significant for males. The number of Facebook friends does not affect the probability in any economically significant way, but is statistically significant for males. The number of mutual friends as well as differences in friend counts are statistically significant for both genders, with differences again being generally disliked, whereas overlap in social circles has a positive effect. A German speaking candidate may be less attractive to a female user, but only if the user herself does not speak German. If both speak German, that effect is cancelled or even reversed.

Distance has no effect, possibly because candidates are all relatively close to each other

[15]For candidates below 18, the absolute distance from 18 is measured. Hence a positive coefficient indicates a preference for younger potential partners, while a positive coefficient for candidates older than 18 indicates a preference for older candidates.

to begin with. The sign of the estimated coefficient is negative though, in line with expectations. Last but not least, women become more selective as they continue rating candidates, with the coefficient on the period variable being negative. Males, on the other hand show no such behavior. Note, as only the first 100 decisions are used in this estimation, changes in acceptance rates over time might also reflect belief updates of new entrants about candidate's behavior.

To further decompose preferences into common and individual components, I run a set of random-effects and fixed-effects regressions. I report loglikelihood and Wald χ^2 statistics in Table 2.3. I start out with a baseline random effects model including attractiveness as the sole covariate, which implements a simple common preference model. I then build up to include more common covariates relating to the candidates' attributes and finally show statistics for the full set of covariates including fixed effects, allowing for individual-specific preferences. Note that I lose some individuals in fixed-effects estimation due to no variation in the dependent variable, which makes direct comparison of loglikelihood values across random effects and fixed effects estimation difficult.

Focussing on random-effects specifications, the attractiveness measure by itself already explains a meaningful part of a user's decision. As the specification allows for more common as well as individual preference parameters, loglikelihood values and Wald statistics increase significantly, but the increase is only marginal. That conclusion also holds when comparing fixed effects specifications. Nevertheless, comparing random effects to fixed effects specifications as well as the common preferences specifications to the introduction of pair-specific covariates, there clearly also is an individual component to preferences aside from common factors, with the corresponding likelihood ratio tests rejecting their respective null hypotheses. So although physical attractiveness of a candidate is a strong and common predictor to individual's willingness-to-date, there remains an individual component with a preference towards homogamy across several dimensions.[16]

[16]Furthermore, R^2 statistics for the linear probability model in the appendix are fairly low, suggesting that a large part of the variation in the dependent variable remains unexplained.

In summary, results are consistent with the findings in Hitsch et al. (2010) and similar research (e.g., Rudder, 2014; Belot and Francesconi, 2013; Fisman et al., 2006), providing evidence that assortative patterns are at least in part due to a combination of common and individual preferences. At the same time, it extends previous research by including a crowd-based attractiveness measure capturing the physical attractiveness of a potential partner. This extension proves to be crucial, as it is by far the most relevant factor in individual's willingness-to-date-decisions.

2.5 Characterizing the initial match

Having estimated preference parameters, this section of the paper moves one step forward to analyze how successful males and females are in searching for a mate. Given the individuals' preferences, I can estimate which candidate in each individual's search sequence is their most preferred mate. This mate is ranked first. I then look at the best-ranked matched mate of an individual, using the application's match definition (both user and candidate responding with HI). As mate search is two-sided, individuals are unlikely to be matched with their first-ranked candidate; the question then is how close males and females they get to rank 1.

Achieved outcomes are highly asymmetrical across gender. Measured in ranks, outcomes of females and males differ by a factor of almost 10: The median best rank achieved by a female is 8, whereas that of a male is 79. This section explains how one ends up in such an equilibrium by employing the theoretical framework of the Secretary Problem.

The first subsection introduces a model deriving a rank prediction for each individual given search length, own acceptance rate and attractiveness. The following subsection tests the model's predictions empirically. Finally, the last subsection validates the model's assumptions.

2.5.1 Model

The goal of this section is to introduce a model offering a framework in which to analyze empirical results on outcomes. The framework needs to adequately reflect the features of the mate search problem in general as well as the application's setup in particular.

Throughout this section, the focus will lie on asymmetries in outcomes across gender. Asymmetric outcomes are well known in game theory and the study of stable matchings (Roth and Sotomayor, 1992)[17], with the standard approach being the study of the stability of matchings (Roth and Sotomayor, 1992). Assuming that all preferences are known and there is a static set of candidates on each side, a stable two-sided matching can always be found, but the assumption itself might be unrealistic in many empirical contexts. Rather than relying on these assumptions, I follow the statisticians' approach taken by Eriksson et al. (2008), where agents base their preferences and rank-orderings only on a subset of potential matching candidates. As Eriksson et al. (2007) argue, in such types of situations where only a small portion of preferences will ever be revealed, it does not make sense to speak about the best overall matching — even more so as the number of stable matchings is asymptotically proportional to $e^{-1}n\ln(n)$ and only characteristics of lower and upper bounds of these matchings are known.[18] Researchers should focus on asymmetries in outcomes and agents' search strategies instead. Also, the set of candidates is dynamic rather than stable, with agents leaving the set when they mate, old cohorts exit even if they do not mate and young cohorts enter. Such a setup is appropriate and more realistic in the current setting.

Eriksson et al. (2008)'s model is itself based on the well known "Secretary Problem" from optimal stopping theory[19] — in particular the one-sided optimal rank version of Lindley

[17] A matching is stable if there is no man and no woman who prefer each other to their current match. In the case of multiple stable matchings in D. Gale (1962), there is always one which is optimal for one side (say women), while at the same time being the pessimal outcome allocation for the other side (the men).

[18] Where e is Euler's constant and n the number of candidates on each side (Pittel, 1989).

[19] Ferguson (1989)

(1961) and Chow et al. (1964), and the two-sided extension of Eriksson et al. (2007).[20] Asymmetries can arise endogenously in the model (in addition to exogenous influences such as uneven sex ratios or different costs of choice), even in cases where the game's setup is perfectly symmetric.

Setup

There is a large universe U of potential candidates. Each agent has $N << U$ periods available for dating, exogenoulsy set before the start of the search. In each period r, available mates are randomly matched to each other, where for each individual, the rank order is independently drawn from a uniform distribution. In other words, the rank of the next date relative to the $r-1$ partners already observed is a random variable drawn from a uniform distribution on the set of ranks from 1 to r. In every period, the best-ranked, worst-ranked, or anything inbetween is equally likely to come up. It should be stressed that individuals do not observe the values of the implicit ranks, but can only rank the candidates they have seen, with the set of ranked candidates expanding with every subperiod. As usual in the Secretary Problem (but contrary to typical assumptions in matching theory), I assume that agents do not have a priori knowledge of the distribution of the characteristics that are manifested in the rank; there is no issue of learning the range of attractiveness of the other sex. Therefore, an individual cannot make any informed decision on the first date — only later comparisons will reveal how good or bad the first date really was.

If both agents at a date accept to get mated, they leave the game. For simplicity, it is assumed that each time an agent leaves the game, another agent of the same sex enters immediately. An agent also leaves the game if she remains unmated after her last period, in which case she gets the payoff of the individual-specific outside option, ranked $\nu_w N$ or $\nu_m N$ (where $w \in W$ refers to women, $m \in M$ to men). This outside option reflects the cost

[20]There exists a variety of different outcomes that can be optimized in the Secretary Problem. The classic one-sided version maximizes the probability of getting the best match; I will assume agents minimize ranks, which is equivalent to maximizing utilities.

of staying single (or the cost of finding a partner through alternative channels). A game is called symmetric if $\nu_w = \nu_m$ for all w and m, and asymmetric if $\nu_w \neq \nu_m$.[21]

All agents minimize the expected rank of their mate. Preferences of different agent's are assumed independent, which precludes the possibility that an agent can draw any conclusion from past candidates' decisions about whether or not future candidates will accept him. This is in contrast to other models assuming the opposite polar case of common preferences (all females have identical preferences over males and vice versa). Common preferences give only one stable matching, while independent preferences of individuals are much more likely to lead to asymmetric outcomes. Both types of preference assumptions are strong and likely unrealistic; they should be seen as baseline cases that allow researchers to solve the mate-search problem. However, as the authors point out, allowing for sufficiently independent preferences is key for the emergence of asymmetric equilibria.

For the sake of simplicity, I discuss the model from the viewpoint of a female. Analogous statements hold for males.

Expected final mate rank

Assuming that each agent minimizes the expected rank of her mate among the N partners she would meet if she completed all N periods, it follows from uniformity and independence assumptions that the expected final rank for a mate who is ranked ρ among the r partners observed up to period r after one more date is

$$\frac{\rho}{r+1}(\rho+1) + \frac{r+1-\rho}{r+1}\rho = \frac{r+2}{r+1}\rho, \tag{2.3}$$

where the first term on the left hand side corresponds to the case where the next date is better ranked than current date ρ (with probability $\rho/(r+1)$ the rank of the current date increases to $(\rho+1)$) and the second term corresponds to the case where the new date is worse ranked than ρ (with probabilty $(r+1-\rho)/(r+1)$, the rank of the current date stays at ρ). By repeating this over all periods $r+1, r+2, \ldots, N$ the expected final rank of

[21] This is a slight modification compared to Eriksson et al. (2008) where outside options are gender-specific rather than individual-specific.

the current mate becomes

$$\mathrm{E}[\rho|mate] = \frac{r+2}{r+1} \cdot \frac{r+3}{r+2} \cdots \frac{N+1}{N}\rho = \frac{N+1}{r+1}\rho. \tag{2.4}$$

This holds although the actual set of candidates an individual would meet in remaining periods is not known.

Strategy

A strategy in this two-sided Secretary Problem is a stopping rule that says for each period r whether or not to accept a date of observed rank ρ in this period. Payoffs in the game are defined by the final mate rank, and expected payoffs depend on the strategy profile of all agents. Define R_r^w as the expected final mate rank for a certain individual of sex W entering period r. Agents want to minimize R_1, the expected final mate-rank at the start of the game. The following recurrence governs the expected final mate-rank when a player of sex W enters period r:

$$R_r^w = P[mate] \cdot \frac{N+1}{r+1} \cdot \mathrm{E}[\rho|mate] + (1 - P[mate]) \cdot R_{r+1}^w. \tag{2.5}$$

The first term on the right hand side defines the expected final rank of the current mate given that the agent matches with that mate, whereas the second term defines the expected final rank given the agent continues dating. If one remains not mated after the last period, one obtains the empty mate $\nu^w N$ or $\nu^m N$ for females and males, respectively. Thus

$$R_{N+1}^w = \nu_w N. \tag{2.6}$$

I will assume $0 < \nu_w, \nu_m \leq 1$. In combination with the total number of periods N, the absolute rank of the outside option gets worse the longer one is willing to search, implying a relatively stronger preference to be mated. Having fixed the payoff after the last period given by the outside option, the problem can be solved backwards.

Outcomes

The game is in a steady state if the proprotion of all available females in a given period is constant. Since all available agents of the opposite sex are equally likely to come up at the next date, the probability that a female will be accepted by a male is always the same (and vice versa). Denote these mean probabilities by α^M, α^W, respectively.[22] In equilibrium, every individual in each period optimizes the expected payoff given the steady state.

Let s_r^w be the threshold defining the female strategy in period r, i.e., the agent accepts if the rank she observes in this period is at most s_r^w. Given that she has reached period r, this means the probability that she will accept is s_r^w/r, resulting in a probability to mate of $P[mate] = \alpha^M s_r^w/r$. Given that the female accepts, the expected observed rank of her partner is $\mathrm{E}(\rho|mate) = (s_r^w + 1)/2$. The individual should accept in period r if the expected final mate rank if she mates now is less than or equal to the expected final mate-rank if she does not mate. As shown by Eriksson et al. (2008) in more detail, this leads to the equilibrium condition

$$s_r^w = \left\lfloor \frac{r+1}{N+1} \cdot R_{r+1}^w \right\rfloor, \quad r = 1, \ldots, N \tag{2.7}$$

with boundary condition $R_{n+1}^w = \nu_w N$. The value of u^w determines the threshold in the last period, whereas α^M determines the rate by which the thresholds are lowered in earlier periods.

The recurrence in Equation (2.5) has no closed form solution, but the authors show that for large N, r and s_r^w it can be approximated by

$$\begin{aligned} R_r^w &\approx R_{r+1}^w - \alpha^M \frac{(R_{r+1}^w)^2}{2N} \\ &\approx \frac{2N}{\alpha^M (N + \gamma^w - r)} \end{aligned} \tag{2.8}$$

with $\gamma^w = 2/(\nu_w \alpha^M)$.

The probability of a female accepting in period r is proportional to the expected final mate rank, $s_r^w/r \approx R_r^w/N \approx 2/(\alpha^M(N + \gamma^w - r))$. Increasing search length increases the

[22] Whether a females prior belief about α^M is correct or learned over time is irrelevant.

expected final mate rank, also because the range of possible ranks expands with N. A higher acceptance rate of candidates, on the other hand, improves the rank, as does a lower ν_w, i.e. lower costs of staying single. It is this relationship in Equation (2.8) that I want to investigate empirically.

As Eriksson et al. (2008) demonstrate, asymmetric outcomes across genders are likely to arise, even in symmetric settings with $\nu_w = \nu_m$. The authors further show that α^W is strictly decreasing in α^M. Put differently, the higher the probability that a male candidate is willing to mate, the less likely a woman is willing to mate. This ties into Theorem 1 in Eriksson et al. (2008), deriving that in any equilibrium, the product $\alpha^W \alpha^M$ is a constant approximated by $3/N$.

Also, there is an advantage of being choosy, i.e. having a low overall acceptance rate. According to Equation (2.8), the expected rank of mates for females is inversely proportional to α^M, which in turn, according to Theorem 1 above, is inversely proportional to α^W.[23] Consequently, the expected rank of a female's match is roughly proportional to the female acceptance rate. Thereby, in an equilibrium where females are choosy compared to males (or believed to be choosy by the other side), females end up with on average better mates. Being choosy has previously been connected to better outcomes via other, exogenously determined factors such as the sex ratio, asymmetric process duration or, in the context of biology, differences in the offspring investment between females and males (Rufus A. Johnstone, 1996).

Special cases

The model discussed here is a generalized case of previous models. There are three particular cases that are interesting in light of the empirical results of this paper. I will discuss them briefly here.

First, $\alpha^M = 1$ and $\nu_w = 1$ replicate the one-sided secretary problem in Lindley (1961) and Chow et al. (1964), with candidates always accepting and high costs of staying single

[23] This relationship is also observed on an individual level in Figure 2.4.

for the individuals. The authors in the cited papers show that in this case, the expected final rank converges to a constant of 3.87 as N grows. Equation 2.8 suggests an even lower rank with

$$\begin{aligned} R_1^w &\approx 2N/(N+1) \\ &\approx 2 \end{aligned} \tag{2.9}$$

In other words, as the opposite side accepts every candidate and the outside option remains unattractive, the problem reduces to a one-sided secretary problem with the expected rank converging to a constant.

The second limit case is the opposite case where α^M tends to zero and candidates on the other side become very choosy. In this steady state, females have to take every chance to try to mate with any male better than their outside option, $\nu_w N$, as the probability of a match is very low. In this case, the expected final rank approaches a number proportional to the candidate's outside option

$$R_r^w \to \nu_w N. \tag{2.10}$$

The third special case is the symmetric case with $\nu_w = \nu_m = \nu$ with ν large, i.e. high costs of staying single. The expected rank before the start of the game then yields

$$R_1^w \approx \frac{2}{\sqrt{3}}\sqrt{N}, \tag{2.11}$$

where the expected final ranks are proportional to the square root of search length. This result is close to the one-cohort case derived in Eriksson et al. (2007) where $R_1^w \approx \sqrt{N}$, suggesting a small differing factor when changing from one- to muliple-cohort scenarios. Acceptance rates are symmetric as well, defined as $\alpha_w = \alpha_m = \sqrt{3/N}$, with its product equal to $3/N$ (see Theorem 1 in Eriksson et al., 2008).

2.5.2 Empirical results on best ranks

Using the parameters obtained by preference estimation, I can construct user-specific rankings of each candidate.[24] As individuals usually accumulate more than one match, one can define different ranks as outcomes. At this stage, I choose to look at the best ranked mate of an individual, which based on her HI/BYE decisions is her best option and therefore the rational candidate to pursue further. These matches need not be symmetric, i.e. whereas woman w might be the best ranked match of man m, man m might not be the best ranked match of woman w. I discuss alternative match choices at a later stage.

Based on Equation (2.8), the goal of this analysis is to connect these achieved outcomes to the three measures search length, attractiveness and acceptance rate. Search length N determines the effort an individual is willing to invest in mate search, measured by the number of decisions an individual takes. Attractiveness, measured by the fraction of positive responses an individual gets, puts an individual in a more favorable position in mate selection and can be directly related to the α parameter in Equation (2.8). Finally, the acceptance rate measures selective behavior, with higher rates equivalent to less selective behavior. Bounded between 0 and 1, it serves as a proxy for ν. Both search length and own acceptance rate can also be linked to an individual's outside option νN, with individuals with less attractive outside options willing put more effort into search and behave less selectively.

Whereas for the estimation of preferences, only the first 100 decisions were analysed, the ranking is now assigned over all candidates of a user. The models estimated here only include individuals that have not been active in the app for at least 90 days, having finished their search for a mate (I extend the sample to all users as a robustness check). Using the estimated rank as the dependent variable potentially introduces some measurement error, but coefficients will still be estimated consistently as long as the measurement error is not

[24]Throughout this paper, ranks are predicted ignoring the duration effect. As the estimated coefficient on the duration effect is zero or close to zero, ignoring the effect alltogether has little effect on rankings

correlated with covariates.[25] However, the power of statistical tests might be reduced.

Even without estimating any model, it is clear from unconditional descriptive statistics that there are stark differences in outcomes between gender, with the median best rank of women at 8, compared to 79 for men.[26] More detailed estimation results of a log-log-model are presented in Table 2.4. The sample includes all observations, while robustness checks in the appendix restrict the sample to different minimum search lengths and include still active users in the dataset as the model derives results in a large N environment. Negative coefficients improve ranks, as lower ranks indicate more attractive mates.

Looking at the results for females, it is striking that one cannot reject the null hypothesis of a zero coefficient of search length — not because of high standard errors, but because the point estimate itself is close to zero. In other words, investing more time in their search does not improve nor worsen females' best rank. As expected, more attractive females get better outcomes (an estimated 1.46 percent improvement for every 1 percent increase in attractiveness). Higher own acceptance rates improve the best rank as well.

Men, in contrast, fare worse. For every 1 percent increase in search length, their best rank rises in tandem by an estimated 0.64 percent. In percentile terms (i.e. $rank/N$), there still is an improvement, as the rank grows at a slower rate than search length. Nevertheless, there is a direct cost of searching longer. Being more attractive improves ranks in the male case, too, but own behavior as measured through the acceptance rate does not affect outcomes in any significant way. Finally, I want to point out that the adjusted R^2-statistic for females is relatively low, whereas the same statistic of 0.659 for males is high.

Figure 2.5 plots the best ranks for females and males against search length (females) and the log-product of the acceptance rate and search length (males), thereby connecting results to limit cases in the theoretical framework discussed previously. In the case of females, best ranks converge to a constant proportional to the individual's attractiveness and acceptance

[25]Note that the predicted rank is based directly on attractiveness, and indirectly on the acceptance rate through the individual fixed effect.

[26]Median and phone-based ranks (conditional on having exchanged phone numbers with someone) show similar patterns, with median ranks 175 and 385 for females and 336 and 753 for males, respectively.

rate, approximately mimicking the one-sided limit case of the Secretary Problem. Search length has no effect on outcomes (partial $R^2 = 0.000$). The corresponding partial R^2 for males is 0.488. The high statistic in the male case can be partly explained by the model's assumption that individuals only rank candidates they have actually seen; in other words, search length N has a direct effect on the range of ranks. Consequently, it makes sense that search length explains a substantial part in the variation of best-achieved ranks. At the same time, it is all the more noteworthy that in the case of females, the search length factor does not play a role at all — as in the case of the one-sided limit case of the model.

In the case of males, outcomes can be approximated by the product of the acceptance rate and search length (partial $R^2 = 0.347$), a proxy for the outside option $\nu_m N$ in the model. Here, empirical results suggests that males find themselves much closer to the limit case of the picky candidates, with their outcomes converging to their outside options. This is not true for females, where the corresponding partial R^2 is 0.020. Combined with the male distribution of acceptance rates in Figure 2.3 in the section on preliminary statistics, these results could be rationalized assuming an approximately uniform distribution over ν_m.

Both these results combined suggest an equilibrium with selectively behaving females and, correspondingly, undemanding males. This behavior pattern translates into highly asymmetric rank outcomes.[27] Females approach their optimal outcome, with ranks converging to a constant and relatively low rank of their best-ranked partner irrespective of their search length. Males, on the other hand, approach their pessimal matching; they are matched with candidates whose utilities roughly correspond to their reservation utilities or outside options. The higher their cost of staying single, the longer they are willing to search and the less selective they behave, leaving them with less and less attractive partners. Note that these are not the average ranks of all the candidates a user has matched with; it's the single best rank in his opportunity set.

[27]Combined with uneven sex ratios and differing outside options across gender, asymmetries could even get stronger.

These outcomes are not deterministic. As mentioned previously, there is a multitude of possible equilibria in a setup like this; I only observe one endogenous realization of one equilibrium. One could easily come up with equilibrias that favor men. However, descriptive statistics of other, similarly structured applications like Tinder and previously found empirical patterns in other studies indicate that females generally behave significantly more selectively than males, which will generally affect their outcomes favorably. If anything, the asymmetric results found here are likely to get even more asymmetric as females' cost of staying single is arguably declining and individuals search longer and marry later.

2.5.3 Validating the model

The uniformity assumption

The uniformity assumption assumes that each candidate shown in a new subperiod is as likely to be ranked first, last, or any rank inbetween in the users ordering, which is crucial to forming expectations about final ranks and deciding whether to accept or reject a candidate. As outlined in Section 2.2, the application sorts candidates by a number of factors which may invalidate that assumption. The ordering of candidates is not recorded by the app, and the continuous entering of new and exiting of existing users makes reengineering of the ordering impossible. However, I can use the data and results of Section 2.4 to test the uniformity assumption. In order to do that, I use the preference estimates to predict an individual ranking order of the first 100 decisions for each user. I then look at which ranks appears at what point in the sequence. By averaging over all individuals, I get a probability estimate for each rank in each subperiod.

The results are first shown graphically in Figure 2.6, with one graph for each gender. The horizontal axis shows subperiods r, the vertical axis the probability of a rank $\Pr(R|r)$ in a given subperiod. Each graph plots the probability curve for rank $R = 1$, the (rounded) middle rank $R = r/2$ and the last rank $R = r$ as well as the theoretical uniform distribution. All probability curves are decreasing in subperiods r as the uniform distribution assigns

probability $\Pr(R) = 1/r$ to each rank $R = 1, \ldots, r$.

The application reproduces the uniform distribution very closely. Best-ranked candidates have slightly lower than uniform probabilities, whereas worst-ranked have higher-than uniform probabilities. Middle ranked candidates are very close to the uniform distribution. If any a priori expectation had to be formed, one would have expected the opposite as the applications algorithm prioritizes more attractive (and therefore better-ranked) candidates, which would lead to probabilites higher than predicted for best-ranked candidates in early subperiods (instead of the lower probabilities seen in the graph). As users continuously enter and exit the application and the algorithm also relies on other factors, this ordering does not seem to leave any significant traces.

I further test the uniformity assumption by estimating

$$\mathrm{E}(R_{ir}) = \frac{r+1}{2} \tag{2.12}$$

which results directly from the model's uniformity assumption. Estimating this model in log-log-form should result in a coefficient close to 1 on the period variable $p = r + 1$ and $-\ln(2) = -0.693$ on the constant.[28] Results using both fixed and random effects are displayed in Table 2.5, providing strong evidence that the uniformity assumption can be assumed as given in the application. The table shows estimated coefficients of 1.012 for the period variable and -0.632 for the constant for females and 0.998 and -0.671 for males. Random effects specifications are virtually identical, with coefficients of 1.015 and (females) and 1.004 (males), respectively. In summary, I can conclude that the uniformity assumption is largely fulfilled.

The independence assumption

The model assumes independent rather than common preferences of agents. This offers several advantages. For one, as argued by Eriksson et al. (2008), sufficiently independent preferences may give rise to multiple, asymmetric equilibria that prefer males or females,

[28] $\ln R = \ln r + 1 - \ln 2 = \ln p - \ln 2$

whereas common preferences lead to unique stable matchings with assortative mating. Independent preferences also simplify the model in that there is no consensus on the attractiveness of a candidate, allowing to treat agents of the same sex equally. At the same time, there is no issue whether agents know their own attractiveness beforehand or learn it over time.

That being said, independence of preferences is a strong assumption which should be considered a baseline case. Clearly, there is a common component to preferences as demonstrated by the highly significant effects of the attractiveness measure in preference estimation, which is the average response of users to a candidate. On the other hand, as indicated in Table A.1 in the appendix, there is substantial variation between individuals. The low R^2 statistics in the linear probability model in Table A.2 point in a similar direction, leaving a large fraction of the variation in the outcome variable unexplained. So although independence of preferences clearly is an oversimplification, the assumption of common preferences made in other models appears to be equally strict and unrealistic.

Candidate universe and sex ratio

The model also assumes that there is a candidate pool larger than any of the search lengths an individual may have in order for uneven sex ratios not to have an impact on strategies and outcomes. If the sex ratio constraint was binding, even slight asymmetries in that ratio may exogenously induce additional asymmetries in outcomes unrelated to the endogenously arising imbalances derived in the model.

Figure 2.7 shows that the sex ratio is mostly constant over time, with the number of registered females and males rising in tandem over time. The assumption of a large enough candidate universe itself is largely fulfilled for both genders. More than 99 percent of females rate less than the 11,302 male candidates in the pool (4,170 at the 90th percentile), and less than 7 percent of males rate all the female candidates (5,460 at the 90th percentile).[29]

[29]Note that the number of decisions can actually be higher than the number of candidates due to excluding sexual orientations (bisexuals) or users having been misclassified in sexual orientation, not completed the login process, not definitively having been accepted in the application's vetting process or blacklisted users. All of these users have been dropped from the dataset during the data cleaning process, but may

So although it is possible that a user sifts through all candidates, the vast majority of users never gets to that point. Even if the sex ratio would turn to be relevant, its presumably negative effect for males would be captured in the gender-specific constant in the results in Table 2.4. Comparing these constants in the table does not indicate any negative effects for males, but they could be confounded with other factors entering the coefficient estimate for constant.

Search length N

I use the number of HI/BYE-decisions taken by a user as her search length N. As in the model, the measure ignores the length of the time period it takes a user to make these decisions. The model assumes that the search length N is preset. I make this assumption, too, thereby presupposing that even before entering the application, candidates set themselves an effort level they are willing to put into their mate search or that the investment in the search is determined by factors that are unrelated to the realized outcome (e.g., leisure time). I also only include users that have not been active in the application for at least 90 days, presumably having finished their search.

One should keep in mind that search length may be endogenous. It is a priori unclear what effects endogeneity would have. It is plausible that users unsatisfied with their matched candidates keep on searching for a better match, leading to an upward bias in the search length coefficient. Note that they simultaneously also increase the cost of staying single, as the outside option is proportional to search length.

But it is also plausible that users who get attractive matches want even more of it and continue searching, whereas others with unsatisfying matches give up. In this case, there would be a downward bias in the search length effect. This hypothesis is supported by Table A.9 in the appendix, presenting results of a regression of search length on different user characteristics. In both the female as well as the male case, older, better educated, more attractive and more selective individuals search for longer. Especially with respect

have shown up in the users search sequence.

to attractiveness, if anything, this supports the latter rather than the former bias.

Unique matches

The model (naturally) assumes that matches are unique. In the application, by contrast, this constraint is not enforced. At the same time, individuals are not just maximizing the number of matches — if that were the case, there is little incentive to behave selectively. As more attractive individuals also behave more selectively, this suggest that even if matches are not unique, individuals target a finite number of matched mates.

As search length is constant within a user, I have to restrict the sample to one match per user in order to be able to identify the model. I choose to look at the best-ranked mate a user gets, as the model assumes minimization of the expected rank and the best ranked match should be the pick from the perspective of a rational individual.

Of course, I could have chosen a match by other metrics than by best rank, in particular picking ranks by the number of messages exchanged within a match or focussing on matches that exchange phone numbers. Whichever alternative metric I choose, I would deviate from the pure optimization of minimizing expected ranks and change the model's setup by allowing for interaction between mates through messaging. Interactions will allow feedback and information about the likelihood of a successful outcome, which is precluded in the model. As the aim of this section is to derive results predicted by the model, the most appropriate choice appears to be to use the best-ranked match of each user in estimation. I will turn to alternative measures later.

Backtracking

The model setup excludes backtracking. Although the application excludes backtracking by design as well, users could circumvent this constraint by simply saying yes to all candidates, or saying yes more often. There is only a very limited number of users doing the former and from the perspective of solving the problem of finding a good mate in a reasonable amount of time, unconditionnally saying yes to all candidates is of little use. One cannot examine whether users have lower reservation values than they would have

were they forced to marry their first match and leave the application, but presumably the threshold is lower in the application as the stakes of saying HI are lower.

Increasing cutoffs

The model predicts increasing cutoff threshold s_r^i as a user approaches her final period. In the last period, she is willing to accept anything that is better than her outside option $\nu^W N$. In general, s_r^i is increasing in r.

Whereas in preference estimation the coefficient on the transaction variable was positive only in the case of males, thresholds are steadily rising for both genders when looking at the final periods of each user.[30] Figure 2.8 shows lower bounds of s_r^i for both males and females from a sample of 3,000 randomly drawn users for either gender. Thresholds are derived by predicting (standardized) ranks according to preference parameters (ignoring the time effect) and conditioning on the user accepting a candidate (a HI-decision). The thresholds are then plotted against the remaining periods in a user's search. In both cases, the upward slope indicates rising thresholds (and therefore less selective behavior) as the users' searches draw to a close, as predicted by the model. In the case of males, the slope flattens out towards the end.

Acceptance rates and match probabilities

The distributions of acceptance rates (i.e., the probability a candidate accepts a user) differ markedly across genders as shown in Figure 2.3. Whereas the distribution of females' acceptance rates is concentrated around a low mean of 0.12, the acceptance rates of males are almost uniformly distributed over the unit interval with a mean of 0.51. While the model makes no claim about acceptance distributions, it does derive a decreasing relationship of acceptance rates with respect to search length and attractiveness. Also, Theorem 1 in Eriksson et al. (2008) states that the product of attractiveness and acceptance rate is constant.

[30]Note that in the case of preference estimation, I focus on the first 100 decisions of a user, while I focus on final periods in the graphs that follow. Users might adapt their behavior in earlier stages due to updating beliefs about acceptance rates.

I can confirm these relationships in the data. With respect to search length, acceptance rates are decreasing for both genders: For females, acceptance rates decrease by 3.3 percent for a 1 percent increase in search length. For males, acceptance rates fall by 2.3 percent for the same increase in search length (regressions not shown). Figure 2.4 in preliminary statistics further illustrates the inverse relationship between acceptance rates and attractiveness derived in Theorem 1 in Eriksson et al. (2008). Finally, related to this relationship is the the distribution of the product $\alpha_w \alpha_m$ (the probability of a match) shown in Figure 2.9, predicted to be constant in the model. While these rates are not exactly constant, their distributions certainly are more concentrated than the one-sided acceptance rates.

Filters

Lastly, in order to examine how strictly users constrain their candidate choice set, I look at the age and distance filters that users can set themselves.[31] Figures A.2 and A.3 in the appendix plot candidates' age and distance ranges considered by users, and put them in relation to the users' own attractiveness approximated by a polynomial. Broadly speaking, although more attractive users are generally more selective in their decisions (see Table 3.1), there seems to be little evidence that this already the case when setting search filters. There is a slight downward trend with increasing attractiveness in all cases except for the distance range of females.

2.6 Match progression

This section focuses on the third stage of the application, with matched individuals exchanging messages. As I will show in this section, matches are far from unique, and limiting analysis on best ranks would be restrictive — especially because the best-ranked candidate from the first step need not necessarily turn out to be the most promising match. I look at who contacts whom, who replies and which pairs exchange phone numbers. These

[31] Note that the application's algorithm may override the distance filter in case too few candidates fulfill the filter's criteria.

outcomes are close to the measures used in Hitsch et al. (2010), with the difference that their first step (contacting a mate) is a later-stage decision in my setup, where users already received a positive signal of mutual interest. The ultimate goal is to connedt elicited preferences from the first stage to final outcomes (i.e., the most promising matches of a candidate).

The first subsection discusses opportunity sets, followed by results on final outcomes.

2.6.1 Opportunity sets

Table 2.6 gives an overview on users and matches, putting the number of decisions into context with the number of variously defined matches. I will call the set of matched candidates the opportunity set. Note that the table is based on data of all users, including those not getting a match.

The average female takes 1,695 HI/BYE-decisions, compared to 2,032 for males. 84 percent of females have at least one match, compared to 72 percent for males. The average number of matches (where a match here is defined as both users saying HI to each other) is significantly higher than one, averaging at 36.63 and 20.67, respectively. In both cases, there is considerable variation around those means. Besides the high variation, the distribution of these variables are also skewed to the right, with median numbers considerably lower than averages.

In what follows in the bottom section of the table, I gradually introduce stricter definitions of matches, based on the number of messages the users exchange (at least one, more than 1, more than 10) as well as whether at least one phone number was exchanged in the chat. An exchanged phone number is interpreted as the strongest signal, as typically users interested in each other will at some point exchange phone numbers and move their exchange to another platform or meet in person. The number of matches with users exchanging many messages and phone numbers is fairly low, in many cases identifying the most promising match of an individual.

When compared to the statistics in Table 3.1, user and candidate attributes reflect the changes that would be expected given the assortative tendencies reported in the previous section (not shown). This is reassuring, as the estimation of preferences only relied on the first 100 decisions of users, whereas the matching dataset comprises all matches of all individuals in the application. In particular, the matching dataset contains relatively more attractive and less selective users (the mean of the standardized measure is above zero) with smaller differences in age. Users in matches are also generally slightly better educated and more sociable (as measured by the number of Facebook friends).

I next turn to opportunity sets. Different from the initial choice set, the opportunity set of woman w, M_w, is the set of men m weakly preferring woman w, that is,

$$m \in M_w \quad \text{if and only if} \quad U_M(w,m) \geq c(m,r)$$

Similarly, a man m's opportunity set is defined as W_m with

$$w \in W_m \quad \text{if and only if} \quad U_W(m,w) \geq c(w,r)$$

As derived by Menzel (2015), the size of the opportunity set grows at the rate of $\sqrt{N}$ for large N. I assess this result empirically in Table 2.7 where the size of the opportunity set, measured as the number of matches, grows at a rate proportional to $\approx N^{0.58}$ for both genders in a simple univariate model. In other words, while best achieved ranks diverge strongly across gender, the size of the set of matches grows at comparable rates. If the specification is expanded to include both the attractiveness measure as well as the acceptance rate of an individual, set growth for females is even higher, with a 0.8 percent expansion for every 1 percent increase in search length. For males, the effect of search length remains unchanged. In both the female and the male case, the size of the opportunity set is positively linked to attractiveness and acceptance rate.

Going beyond just the size of the opportunity set, one can further look at inclusive values, typically used for welfare analysis in the context of conditional logit models. Rather

than being a measure of the final outcome, it characterizes an individual's indirect utility derived from having access to a given opportunity set. The inclusive value is defined as the conditional expectation of a woman w's indirect utility function from a choice set M,

$$\mathrm{E}\left[\max_{m\in M\cup 0} U_W | x_w, x_j, (x_m)_{m\in M}\right] = \ln\left(1 + \sum_{j\in M} \exp\left\{U(x_j, x_w, x_m)\right\}\right) + \kappa \tag{2.13}$$

$$= \ln\left(1 + I_w[M]\right) + \kappa \tag{2.14}$$

where the set includes the outside option of staying single denoted by a zero, $I_w[M] = \frac{1}{n^{1/2}} \sum_{m\in M} \exp\left\{U(x_w, x_m)\right\}$, and κ is Euler's constant (Menzel, 2015; McFadden, 1973). Inclusive values grow with both the size of the opportunity set (the number of components of the sum) as well as the quality of potential partners, reflected in $U(x_j, x_w, x_m)$. The relationship between the inclusive value $I_w[M]$ and expected indirect utility gives inclusive values a straightforward interpretation as a surplus measure that can be used for welfare analysis, and can be seen as the indirect utility an individual gets from an expanded choice set. In very general terms, if the choice set is expanded by an alternative better than the best previous alternative, it is considered a welfare improvement.

I compute the inclusive values as defined above by using the estimated $x'b$ indices from preference estimation. In order to take the sequentiality of mate search into account, I compute a second, "chronological" inclusive value that ignores all new matches that are worse than the best of all previously collected matches in the individual's opportunity set. Distributions of both measures grouped by gender are depicted in Figure 2.10. Women generally fare better than men, with female opportunity sets stochastically dominating their male counterparts while at the same time exhibiting lower variance. This is true for both measures.

However, comparability across genders of these indirect utilities is restricted as these calculations are based on (gender-specific) $x'b$ indices. Therefore, I also display inclusive values based on ranks instead of the $x'b$ index, depicted in Figure 2.11.[32] The conclusion

[32]Rank-inclusive values are defined as $\sum_{j\in M} \exp(-rank_j)$.

remains the same — not just the minimum achieved rank is lower for females, but their entire opportunity set is more attractive.

2.6.2 First impressions and final matches

Picking the best-ranked candidate as the final match is a reasonable choice in the context of the first stage and the two-sided Secretary Problem, where individuals minimize ranks based on limited information on the candidate. The best-ranked candidate in the first stage need not be the most promising match in the longer term, however.

This subsection looks at the third stage, where matched individuals are allowed to interact and exchange information via chat messages. By gaining additional information, individuals might choose to deviate from their best-ranked mates; the goal of this section is to analyze such deviations. Instead of assuming the best-ranked match according to the decisions of the first stage is also the match pursued in the longer term, I define final matches by stronger signals of interest. Specifically, with most matched candidates never exchanging any messages, I look at which matched individuals start a conversation and by how much their decision to start exchanging messages is influenced by the assigned first stage ranking of the candidate. Next, I look at whether the individual who started a conversation gets a reply. I then move to an even more conservative match definition, where a pair is seen as a match whenever they exchange at least one phone number. Exchanging a phone number is a strong signal of interest, and typically leads to the pair leaving the application and continuing their exchange elsewhere.

Estimation

As in the first stage, individuals maximize utilities in a discrete choice framework. The utility of match j with man m from the perspective of a woman w is defined as

$$U_W(X_m, X_w; \alpha_W, \gamma_W) = x' b_{wm} \alpha + z'_{wm} \gamma + \epsilon_{wm} \tag{2.15}$$

where $x' b_{wm}$ is the index from the first stage preference estimation, proxying the first

impression, z_{wm} is a vector containing the number of sent and received messages within a match, and ϵ_{wm} is an idiosyncratic error term following a standard logistic distribution. I replace the first impression index $x'b_{wm}$ by rank and percentile measures in different specifications of the model, where ranks and percentiles are defined both over the whole set of candidates as well as over the opportunity set. Utilities for males are defined analogously. Note that as I estimate the model for females and males separately, I avoid any issues concerning within-match correlation of error terms.

Results

As before, analysis is restricted to individuals who have been inactive in the application for at least 90 days. Table 2.8 presents results of fixed effects logit estimations for females for three different dependent variables: A dummy indicating that a user initiated a conversation[33], a dummy indicating that an individual replied to an initiated conversation[34], and a dummy indicating the exchange of a phone number[35]. In the first specification, outcomes are regressed on the own $x'b$ index and the corresponding index assigned by the candidate to the user. In the case of the phone dummy, the number of sent and received messages is included as well. The second specification swaps indices for ranks and adds the squared difference in ranks[36]. Finally, the third specification uses rank percentiles instead of absolute ranks as covariates. Ranks and percentiles are once defined over all decisions taken by a user, once only within the opportunity set. By construction, results for males are identical but mirrored as the user-candidate roles are swapped. I therefore omit them here; results for males can be found in Table A.10 in the appendix. Slight differences in the number of observations are due to bisexual candidates, users that may have not fully passed the initial screening test or have been blacklisted later or due to no variation in the dependent variable.

Results across the xb index, the rank or the rank percentile specification are comparable.

[33]Mean values: 0.098 (females), 0.343 (males).
[34]Mean values (conditional on being contacted): 0.238 (females), 0.088 (males).
[35]Mean values: 0.035 (females), 0.031 (males).
[36]The difference is squared to avoid multicollinearty issues.

The ranking of a candidate has a positive effect on the probability of starting a conversation, with positive coefficients on the index variable and negative for ranks and percentiles (the minimal rank being the most attractive candidate). In general, this means that snap judgments in the first stage are aligned with decisions at later stages, even when conditioned on a smaller, more homogeneous set of candidates. Rank differences within a pair do not seem to play a role once the ranks of both user and candidate are controlled for.

In the case of males, the ranking the candidate gives to the user is either ignored by the party taking the action or even has a negative effect, suggesting that those users taking the initiative aim high, with the consequence that mates get primarily contacted by matched partners they assign lower ranks to. To some extent, this may help decrease the asymmetries found in the previous analysis based on best ranks: Overall, median ranks of partners with whom an individual exchanges a phone number are still asymmetric across gender, but relatively more balanced. The median rank for females is 144, while for males it is 416.5 — shrinking the difference across gender from almost 10 in best ranks to less than 4 in the phone-based match definition[37]. Mean values are 378 and 676, respectively. Females, on the other hand, are more likely to contact a male if the female herself is higher in the male's ordering.

Effects on reply probabilities for females are comparable to the effects on starting a conversations, with first stage preferences translating consistently into the second stage. In the case of males, there are only few significant effects. The ones that are significant point in the same direction as for females. Overall then, the factors affecting both starting a conversation as well as replying to a first contact are similar, but in terms of who makes these steps, roles are clearly assigned: males reach out, females reply.

Finally, whether two individuals exchange a phonenumber in a chat is not influenced by a female's rank ordering, but is affected by the ranking of a male. Aside from the ranking, interactions as measured by exchanged messages play an important role. Estimated coeffi-

[37]The calculation of the above median rank values take the lower rank in case an individual exchanges phone numbers with more than one matched partner.

cients on both the number of sent as well as the number of received messages are positive, consistent with the hypothesis that stronger mutual interest in a match goes along with more exchanged messages, ultimately leading to the exchange of a phone number.

2.7 Conclusion

In summary, this paper looks at three aspects of mate search. First, I analyze binary willingness-to-date decisions of individuals with a (largely) exogenously imposed search sequence to reveal their preferences. Results show a decisive role for the attractiveness of a candidate, where attractiveness is measured by the overall ratio of positive responses a candidate gets. Assuming the measure mostly captures the content of photos, the result confirms previous research such as Rudder (2014) while at the same time extending other studies such as Hitsch et al. (2010), which only had limited access to such measures. Results also show tendencies towards homogamy, with individuals preferring partners similiar to them across several dimensions. Even though there is some evidence for strategic behavior, controlling for such behavior does not alter the estimated preference parameters.

In a second step, I use the estimated preference parameters to construct individual rank orderings and analyze behavior and outcomes in the theoretical framework of a two-sided Secretary Problem. Unlike other matching models, the Secretary Problem as set up by Eriksson et al. (2008) only has to assume preferences over the subset of candidates a user has actually seen, an advantage over other models.

Males and females show stark differences in behavior that show up similarly in descriptive statistics of other applications, which makes it plausible that the differences in my data are not just an outlier. These differences in behavior in turn contribute to asymmetries in outcomes. Whereas the best-achieved candidate rank of females converges to a constant, ranks grow with search length in the case of males. I argue that females and males are close to facing two different limit cases of the theoretical model: The results for females sug-

gest an almost one-sided problem, whereas males face such selective candidates that their outcomes converge to their respective outside options, proxied by their search length and own acceptance rate. While efficient, such asymmetric outcomes might not be desirable. Translated into the marriage market, the findings suggest that existing asymmetries as in the marriage squeeze or due to uneven sex ratios may get worse as couples marry later in life, as marrying later can be seen as extended search length. Previous research also has shown that asymmetries within pairs lead to higher breakup rates.

In a third step, I look at the later stage where individuals are interacting. Taking their preference index, ranks and percentile ranks for candidates from their first stage decisions, I test how this initial decision with limited information relates to later-stage signals of interest. Again, I find asymmetries, with males overshooting in first contact actions, whereas females approach males where both partners are attractively ranked. Overall, later-stage decisions are largely consistent with first-stage decisions across both gender groups.

In conclusion, these findings contribute to the literature on assortative mating, but also offer insights into search behavior and asymmetries in outcomes, which traditionally have been hard to track empirically. It is easy to draw parallels to other settings, most notably job search (Autor, 2001). As in the mate search problem, selective job recruiters face a nearly infinite pool of submitted resumes of candidates, candidates choose among a multitude of job advertisements, and both companies and workers differ in their attractiveness and selectivity.

References

Autor, D. H. (2001). Wiring the labor market. *The Journal of Economic Perspectives*, 15(1):25–40.

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political Economy*, 81(4):813–846.

Belot, M. and Francesconi, M. (2013). Dating preferences and meeting opportunities in mate choice decisions. *Journal of Human Resources*, 48(2):474–508.

Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the us income distribution. *European Economic Review*, 43(4):853–865.

Chiappori, P.-A. and Salanie, B. (2016). The econometrics of matching models. *Journal of Economic Literature*.

Chow, Y., Moriguti, S., Robbins, H., and Samuels, S. (1964). Optimal selection based on relative rank (the "secretary problem"). *Israel Journal of Mathematics*, 2(2):81–90.

D. Gale, L. S. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

Eriksson, K., Sjöstrand, J., and Strimling, P. (2007). Optimal expected rank in a two-sided secretary problem. *Operations Research*, 55(5):921–931.

Eriksson, K., Sjöstrand, J., and Strimling, P. (2008). Asymmetric equilibria in dynamic two-sided matching markets with independent preferences. *International Journal of Game Theory*, 36(3-4):421–440.

Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science*, 4(3):282–289.

Fisman, R., Iyengar, S. S., Kamenica, E., and Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, pages 673–697.

Gray, D., Yu, K., Xu, W., and Gong, Y. (2010). Predicting facial beauty without landmarks. In *European Conference on Computer Vision*, pages 434–447. Springer.

Guven, C., Senik, C., and Stichnoth, H. (2012). You can't be happier than your wife. Happiness gaps and divorce. *Journal of Economic Behavior & Organization*, 82(1):110–130.

Hitsch, G. J., Hortaçsu, A., and Ariely, D. (2010). Matching and sorting in online dating. *American Economic Review*, 100(1):130–63.

Kashyap, R., Esteve, A., and García-Román, J. (2015). Potential (mis) match? Marriage markets amidst sociodemographic change in India, 2005–2050. *Demography*, 52(1):183–208.

Lee, S. (2016). Effect of online dating on assortative mating: Evidence from South Korea. *Journal of Applied Econometrics*, 31(6): 1120–1139.

Lindley, D. V. (1961). Dynamic programming and decision theory. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 10(1):39–51.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142.

Menzel, K. (2015). Large matching markets as two-sided demand systems. *Econometrica*, 83(3):897–941.

Pittel, B. (1989). The average number of stable matchings. *SIAM Journal on Discrete Mathematics*, 2(4):530–549.

Rosenfeld, M. J. and Thomas, R. J. (2012). Searching for a mate the rise of the internet as a social intermediary. *American Sociological Review*, 77(4):523–547.

Roth, A. E. and Sotomayor, M. (1992). Chapter 16: Two-sided matching. Volume 1 of *Handbook of Game Theory with Economic Applications*, pages 485 – 541. Elsevier.

Rothe, R., Timofte, R., and Van Gool, L. (2015). Some like it hot — visual guidance for preference prediction. *arXiv preprint arXiv:1510.07867*.

Rudder, C. (2014). *Dataclysm: Love, Sex, Race, and Identity–What Our Online Lives Tell Us about Our Offline Selves*. Crown.

Rufus A. Johnstone, John D. Reynolds, J. C. D. (1996). Mutual mate choice and sex differences in choosiness. *Evolution*, 50(4):1382–1391.

Sales, N. J. (2015). Tinder and the dawn of the dating apocalypse. *Vanity Fair*.

Smith, A. (2016). 15% of american adults have used online dating sites or mobile dating apps. *Pew Research Center*.

Willis, J. and Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17(7):592–598.

Table 2.1: Summary statistics on user-candidate attributes in decisions

	Female		Male	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*
HI	0.141	0.348	0.498	0.500
Attractiveness	0.487	0.164	0.072	0.071
Attractiveness, *standardized*	0.011	1.000	0.007	1.000
Squared diff. in attractiveness, *positive*	0.237	0.629	0.874	3.333
Squared diff. in attractiveness, *negative*	2.001	4.591	1.003	1.632
Acceptance rate	0.102	0.104	0.489	0.291
Acceptance rate, *standardized*	-0.087	0.863	-0.041	0.987
Age	25.73	4.938	26.62	5.135
Squared diff. in age, *positive*	4.096	14.61	19.47	36.56
Squared diff. in age, *negative*	15.50	25.45	7.155	20.90
University	0.069	0.253	0.087	0.282
Both university	0.009	0.093	0.007	0.085
Same school	0.088	0.283	0.076	0.266
German speaking	0.843	0.364	0.752	0.432
Both German speaking	0.648	0.478	0.640	0.480
No. of friends, *in hundreds*	4.887	3.485	5.863	4.513
Mutual friends	0.261	2.316	0.231	2.268
Squared diff. in friends, *positive*	0.903	7.460	2.060	10.82
Squared diff. in friends, *negative*	3.237	13.959	1.409	9.275
Distance in *km*	41.64	50.08	39.69	52.41
TRX	87.13	11.36	89.20	13.85
% of search covered	0.083	0.135	0.080	0.139
Observations	453,575		821,525	

Source: BLINQ; own calculations. Based on the 100 first decisions of all users. Note that means and standard deviations refer to decisions, not users, thereby giving more weight to more active users.

HI is a dummy indicating a positive decision. *Attractiveness* is defined as the ratio of the number of *HI*'s a user got, divided by the number of times the user has been rated. The measure is standardized within gender. Differences are taken over the standardized measure. *Acceptance rate* is defined as the ratio between the number of times a user rates *HI*, divided by the total number of decisions she has taken. The user's own decision has been calculated out of the attractiveness measure. Standardization as in the case of attractiveness. *Age* is measured in years and bounded between 13 and 40. *University* is a dummy indicating whether the user has a university listed on his Facebook profile. *Both university* is a dummy indicating whether $university == 1$ for both the user as well as the candidate.*Same school* is a dummy indicating whether both user and candidate list the same school on their Facebook profile. *German speaking* is a dummy indicating the language set in the app. *No. of friends* is the number of Facebook friends measured in hundreds, *mutual friends* the number of mutual friends that also use the dating application. *Squared difference in friends* is measured in units of 100,000. *Distance* is the distance in *km* between user and candidate, where the information on location was drawn just once, assuming users do not move. Only candidates within a $300km$ radius are considered. *TRX* is the number of decisions a user has taken, capped at 100. *% of search covered* is the fraction of the current decision number divided by the total number of decisions taken by a user.

Table 2.2: Fixed effects logit results on preference estimates (coefficients)

	Female		Male	
	Coeff.	*SE*	*Coeff.*	*SE*
DV: Binary willingness-to-date decision				
Attractiveness, *standardized*	0.892	0.010	1.322	0.008
Squared diff. in attractiveness, *positive*	-0.002	0.017	-0.040	0.004
Squared diff. in attractiveness, *negative*	-0.067	0.002	-0.117	0.004
Years ≥ 18, *absolute*	0.050	0.004	-0.042	0.002
Years < 18, *absolute*	-0.609	0.042	0.174	0.026
Squared diff. in age, *positive*	-0.014	0.001	-0.006	0.000
Squared diff. in age, *negative*	-0.014	0.001	-0.001	0.000
University	0.037	0.016	-0.018	0.012
Both university	0.052	0.057	0.067	0.038
Same school	0.022	0.020	0.031	0.013
German speaking candidate	-0.089	0.028	-0.014	0.016
Both German speaking	0.084	0.031	0.039	0.018
No. of friends, *in hundreds*	0.001	0.002	0.004	0.001
Mutual friends	0.031	0.002	0.004	0.001
Squared diff. in friends, *positive*	-0.004	0.002	-0.004	0.002
Squared diff. in friends, *negative*	-0.002	0.001	-0.003	0.001
Distance in *km*	-0.000	0.000	-0.000	0.000
TRX	-0.004	0.000	0.000	0.000
Observations	441,790		785,936	
Individuals	5,367		9,550	
Log-likelihood	-129,430		-319,096	

Source: BLINQ; own calculations. Based on the 100 first decisions of all users. For females, 368 users (11,785 observations) were dropped because of all positive or all negative outcomes. For males, 947 users (35,589 observations) were dropped because of all positive or all negative outcomes.
Variable definitions as before.
Variables other than direct user-candidate comparisons relate to the candidate, not the user. User characteristics are captured in the fixed effect.

Table 2.3: Common vs individual preferences decomposition

	Female		Male	
	Loglikelihood	*Wald* χ^2	*Loglikelihood*	*Wald* χ^2
Candidate attractiveness, RE	-150,985	27,666	-364,286	87,731
Candidate attractiveness, FE	-132,191	29,316	-320,887	111,249
Common preferences, RE	-150,900	27,748	-364,116	87,977
Common preferences, FE	-131,749	30,200	-320,697	111,629
Full set of covariates, RE	-148,923	29,774	-362,633	88,549
Full set of covariates, FE	-129,430	34,839	-319,096	114,831
Observations, RE	453,575		821,525	
Individuals, RE	5,735		10,497	
Observations, FE	441,790		785,936	
Individuals, FE	5,367		9,550	

Source: BLINQ; own calculations. Based on the 100 first decisions of all users.

For females, 368 users (11,785 observations) were dropped in fixed effects estimation because of all positive or all negative outcomes. For males, 947 users (35,589 observations) were dropped because of all positive or all negative outcomes.

Random effects specifications assume a normally distributed random effect.

Candidate attractiveness includes a candidate's attractiveness measure as the sole covariate. *Common preferences* includes covariates relating to the candidate; specifically *attractiveness*, *age*, *university education*, *Facebook friend count*, and *language*. *Full set of covariates* additionally includes differences in *attractiveness*, *age*, *university education*, *Facebook friends* between individual and candidate and controls for *distance*, whether they both went to the *same school*, *mutual friends*. All specifications include a control for the decision number to take potential duration effects into account.

Full results on the fixed effect estimation are reported in Table 2.2.

Table 2.4: Best achieved rank explained by search length, attractiveness and acceptance rate

	Female		*Male*		*Female*		*Male*	
	Coeff.	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*
DV: $\ln bestrank$								
$\ln N$	0.084	(0.021)	0.584	(0.015)	-0.002	(0.020)	0.643	(0.012)
$\ln attract$					-1.461	(0.071)	-1.152	(0.021)
$\ln accrate$					-0.543	(0.032)	-0.026	(0.021)
Constant	1.515	(0.131)	0.484	(0.101)	-0.591	(0.151)	-3.417	(0.121)
Observations	2,652		3,381		2,652		3,381	
R^2	0.001		0.272		0.183		0.659	
F-Stat	14.96		1,262		202.2		1,728	

Source: BLINQ; own calculations.

The sample considers the best-ranked matched mate for users who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match. The dependent variable *lnrank* is the logarithm of the individual-specific rank of the matched mate, where the rank is based on the estimated preference parameters reported previously. $\ln N$ is the logarithm of the length of an individual's search sequence, i.e. the number of decisions a user has taken. $\ln attract$ and $\ln accrate$ are the logarithmized measures of *attractiveness* and *accrate* reported previously.

Table 2.5: Testing the uniformity assumption

	Female		Male	
	FE	*RE*	*FE*	*RE*
DV: $\ln R$				
$\ln p$	1.012	1.015	0.998	1.004
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	-0.632	-0.643	-0.671	-0.695
	(0.004)	(0.004)	(0.003)	(0.003)
Observations	453,575	453,575	821,525	821,525
Individuals	5,735	5,735	10,497	10,497
Fixed Effects	*Yes*	*No*	*Yes*	*No*
R^2	0.650	0.650	0.584	0.584

Source: BLINQ; own calculations. Standard errors in parentheses.

Table 2.6: Choices and matches

	Female		Male	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*
Individuals	6,066		11,302	
in %	*0.349*		*0.651*	
No. of decisions taken	1,695	1,836	2,032	2,126
No. of HI's	141.6	235.2	969.0	1,316
as a percentage of decisions	*0.116*	*0.126*	*0.506*	*0.296*
Prob. of at least 1 match	0.839	0.367	0.724	0.447
No. of matches	36.63	61.17	20.67	41.63
as a percentage of decisions	*0.033*	*0.058*	*0.013*	*0.036*
No. of matches, message exchanged	11.33	20.06	8.372	18.43
as a percentage of decisions	*0.009*	*0.019*	*0.005*	*0.015*
as a percentage of matches	*0.306*	*0.212*	*0.436*	*0.315*
No. of matches, >1 message	8.333	15.05	5.528	13.09
as a percentage of decisions	*0.008*	*0.015*	*0.004*	*0.009*
as a percentage of matches	*0.220*	*0.174*	*0.266*	*0.248*
No. of matches, >10 messages	2.936	5.973	1.881	5.167
as a percentage of decisions	*0.003*	*0.006*	*0.001*	*0.003*
as a percentage of matches	*0.077*	*0.099*	*0.085*	*0.136*
No. of matches, phoneno. exchanged	0.655	1.812	0.465	1.609
as a percentage of decisions	*0.001*	*0.002*	*0.000*	*0.001*
as a percentage of matches	*0.017*	*0.040*	*0.021*	*0.065*

Source: BLINQ; own calculations. Based on all users in the database.

Table 2.7: Results on the expansion of the opportunity set (all observations)

	Female				*Male*			
	Coeff.	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*
DV: $\ln totalmatches$								
$\ln N$	0.583	(0.014)	0.801	(0.008)	0.580	(0.013)	0.572	(0.011)
$\ln attract$			0.928	(0.032)			0.934	(0.022)
$\ln select$			0.939	(0.012)			0.578	(0.018)
Constant	-1.130	(0.093)	0.732	(0.067)	-1.864	(0.084)	1.587	(0.121)
Observations	2,652		2,652		3,381		3,381	
R^2	0.373		0.828		0.351		0.669	
F-Stat	1,717		4,137		1,986		1,945	

Source: BLINQ; own calculations. The size of the opportunity set is measured as the total number of matches of a user. One observation per individual.

The sample considers users who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match. The dependent variable *lnmatches* is the logarithm of the individual-specific total number of matches. $\ln N$ is the logarithm of the length of an individual's search sequence, i.e., the number of decisions a user has taken.

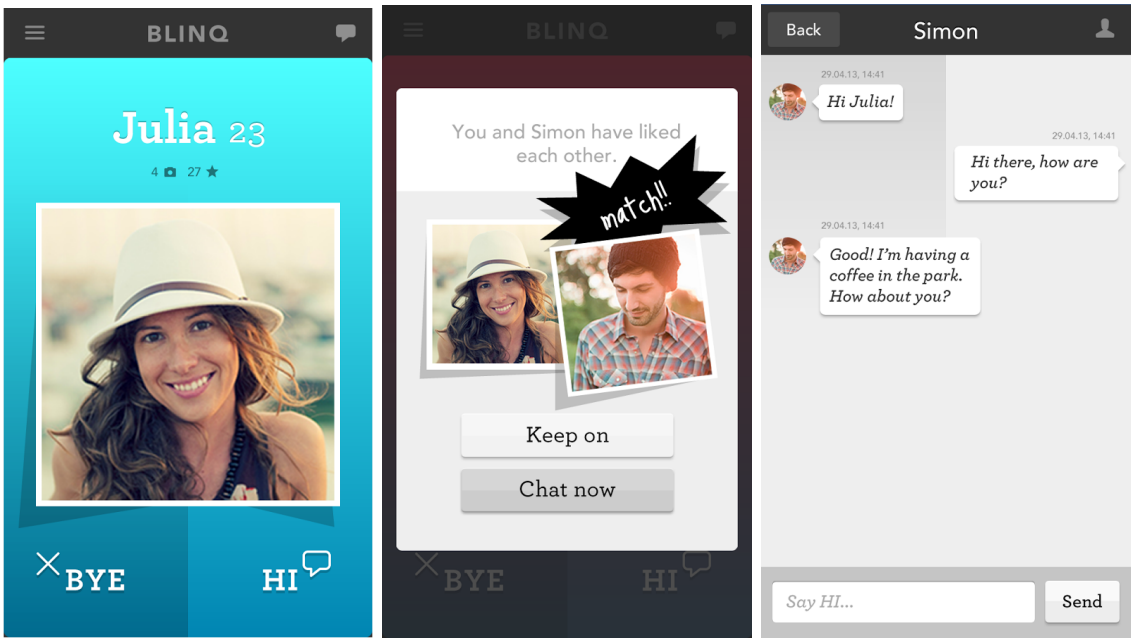
Attractiveness and *acceptance rate* are standardized within gender as previously.

As shown by Menzel (2015), the size of the oppurtunity sets grows as $\sqrt{N}$, which implies a coefficient of 0.5 on $\ln N$.

Table 2.8: First impressions in later stages, females

	Measures based on search length			Measures based on matches		
	Convstart	*Reply*	*Phone*	*Convstart*	*Reply*	*Phone*
Specification 1						
xb1	0.126	0.023	-0.004			
	(0.005)	(0.003)	(0.008)			
xb2	-0.146	0.424	0.186			
	(0.042)	(0.035)	(0.086)			
sentmess			0.070			
			(0.005)			
recmess			0.030			
			(0.004)			
logL	-14,219	-28,924	-5,194			
Observations	54,336	61,894	47,505			
Specification 2						
rank1	-0.055	-0.009	0.002	-0.012	-0.004	0.001
	(0.002)	(0.001)	(0.004)	(0.001)	(0.000)	(0.001)
rank2	-0.004	0.019	0.011	0.003	-0.003	-0.000
	(0.002)	(0.001)	(0.004)	(0.000)	(0.000)	(0.001)
rankdiffsq	0.000	-0.000	-0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sentmess			0.071			0.070
			(0.005)			(0.005)
recmess			0.029			0.030
			(0.004)			(0.004)
logL	-14,281	-28,893	-5,192	-14,154	-28,836	-5,190
Observations	54,336	61,894	47,505	54,336	61,894	47,505
Specification 3						
pctile1	-1.563	-0.168	0.042	-1.415	-0.240	-0.040
	(0.059)	(0.037)	(0.099)	(0.051)	(0.034)	(0.089)
pctile2	-1.697	-1.455	-0.718	0.914	-1.451	-1.222
	(0.175)	(0.141)	(0.350)	(0.091)	(0.060)	(0.160)
sentmess			0.070			0.071
			(0.005)			(0.005)
recmess			0.030			0.028
			(0.004)			(0.004)
logL	-14,114	-28,955	-5,195	-14,089	-28,717	-5,167
Observations	54,336	61,894	47,505	54,336	61,894	47,505

Source: BLINQ; own calculations. Standard errors in parentheses.
The sample considers users who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match. *Convstart* is a dummy variable indicating whether a user starts a conversation, *reply* an indicator whether a user replies to a started conversation (conditional on the conversation being started). *Phone* is a dummy variable indicating whether a phone number was exchanged. $xb1$ and $xb2$ are the indices calculated according to estimated preference parameters for user and candidate, respectively. $rank1$ and $rank2$ are the ranks calculated based on the indices (in hundreds for the left half of the table), with rank1 equal to 1 being the most attractive candidate presented to the user. In the $rankdiffsq$ is the squared difference in ranks, measured in 10,000 units in the case of the left half of the table. $pct1$ and $pct2$ are the respective percentile ranks. *sentmess* is the number of messages sent to the candidate, *recmess* the number of messages received.
Observations number differ because of no variation at the individual level as well as bisexual candidates.



(a) Step 1 (b) Step 2 (c) Step 3

Figure 2.1: Subgame decisions leading to a match

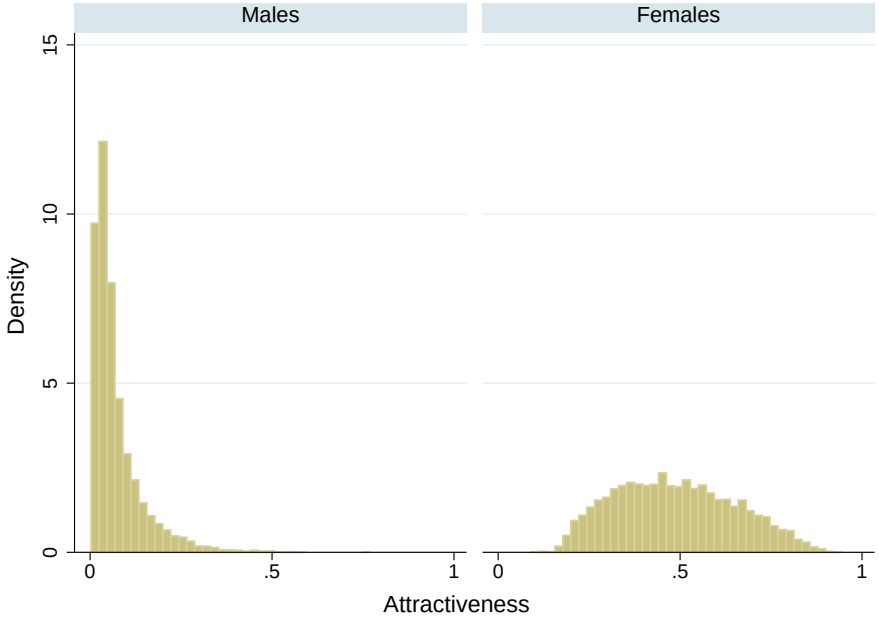


Figure 2.2: Attractiveness $a = \frac{liked}{liked+disliked}$, by gender

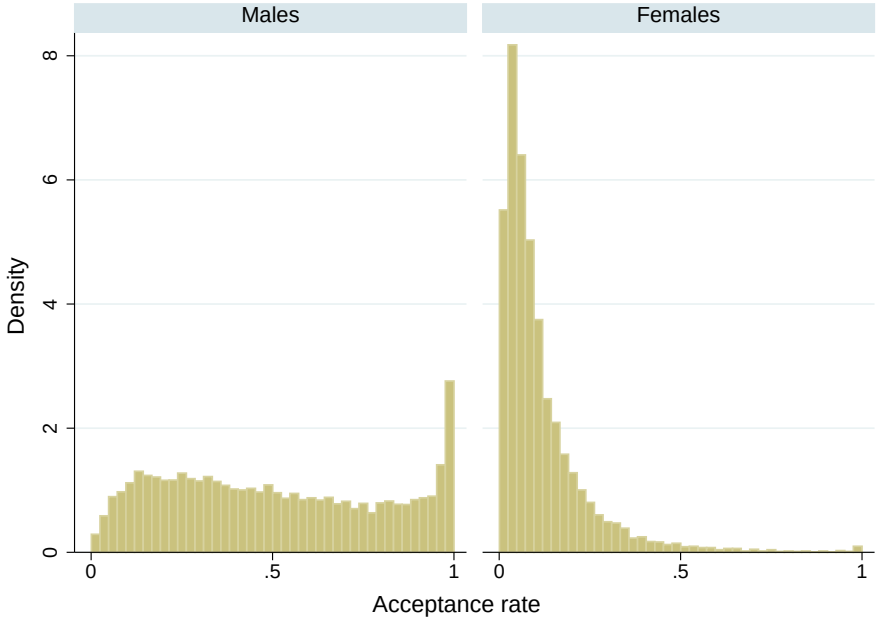
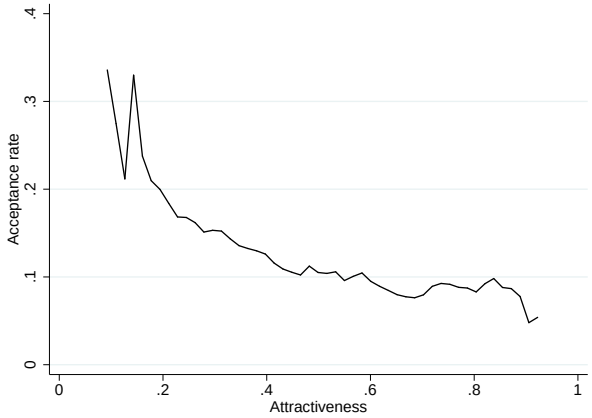
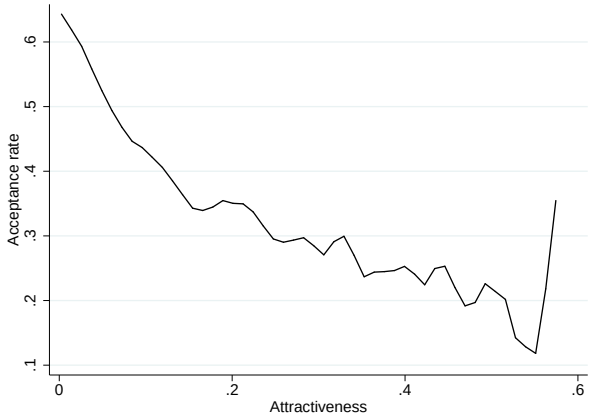


Figure 2.3: Acceptance rate $s = \frac{likes}{likes+dislikes}$, by gender

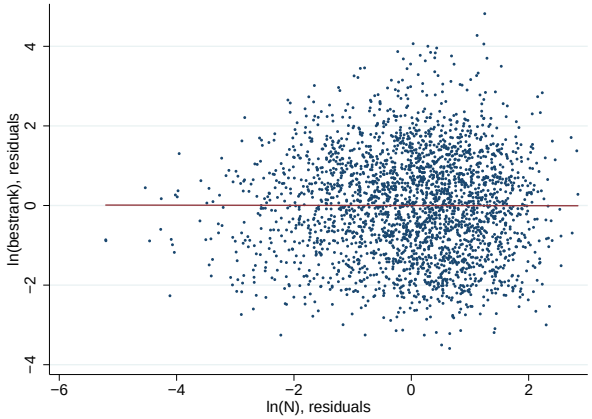


(a) Female

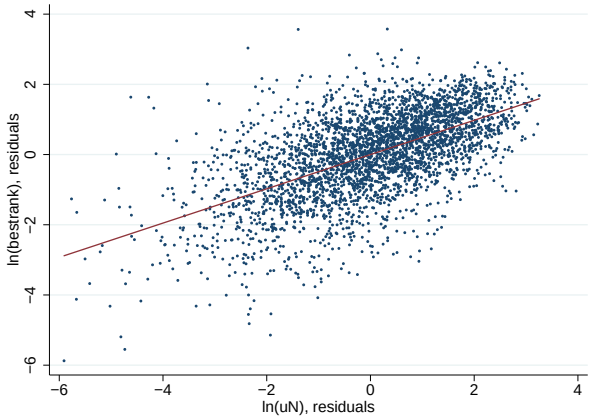


(b) Male

Figure 2.4: Acceptance rates vs attractiveness, by gender

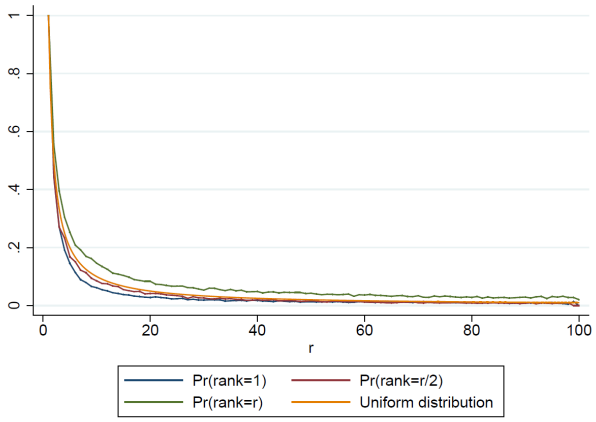


(a) Female

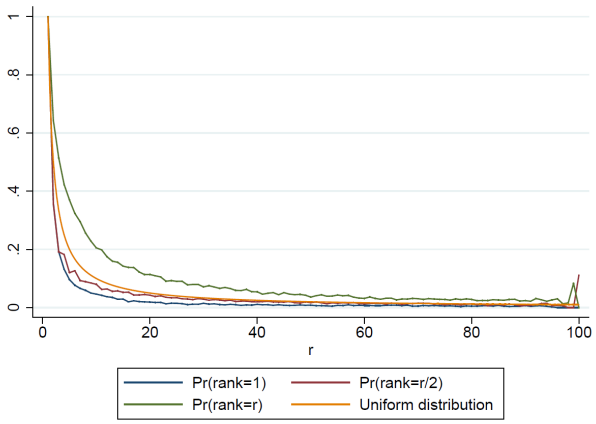


(b) Male

Figure 2.5: Outcomes as limit cases, by gender



(a) Male



(b) Female

Figure 2.6: Probabilities of different ranks across subperiods

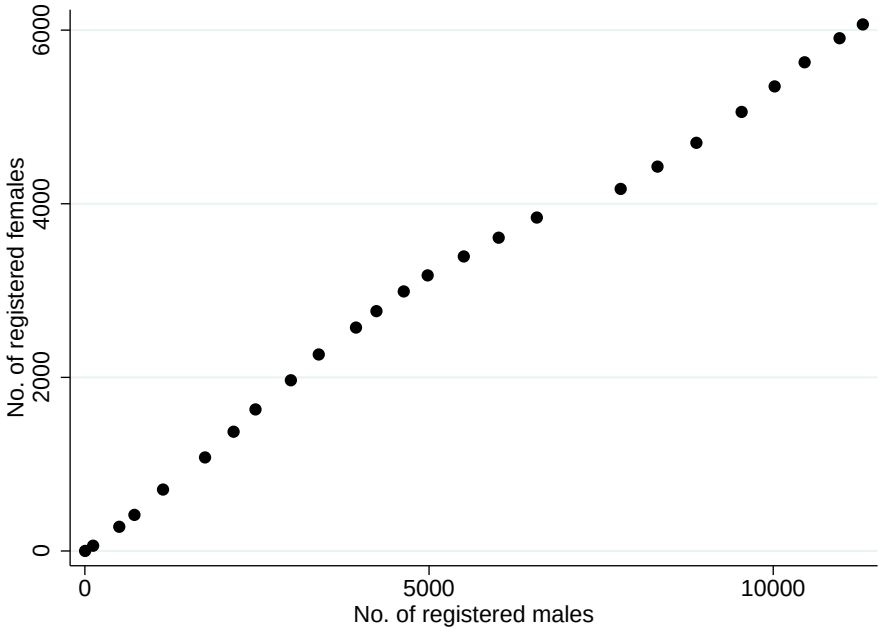
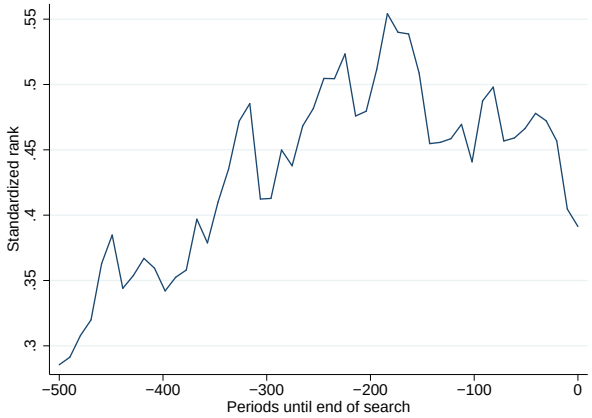
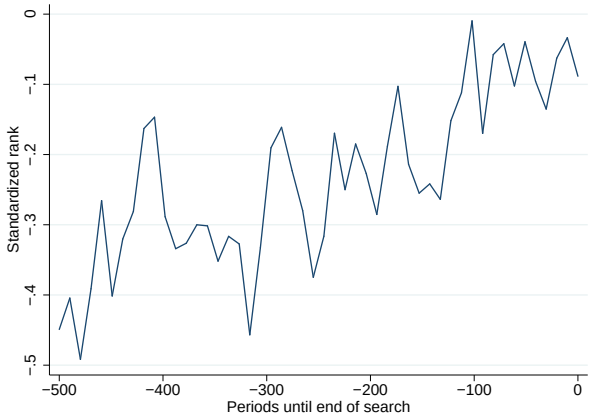


Figure 2.7: Number of registered users, by month



(a) Male



(b) Female

Figure 2.8: s_r lower bound as measured by predicted ranks (conditioned on $Y = HI$)

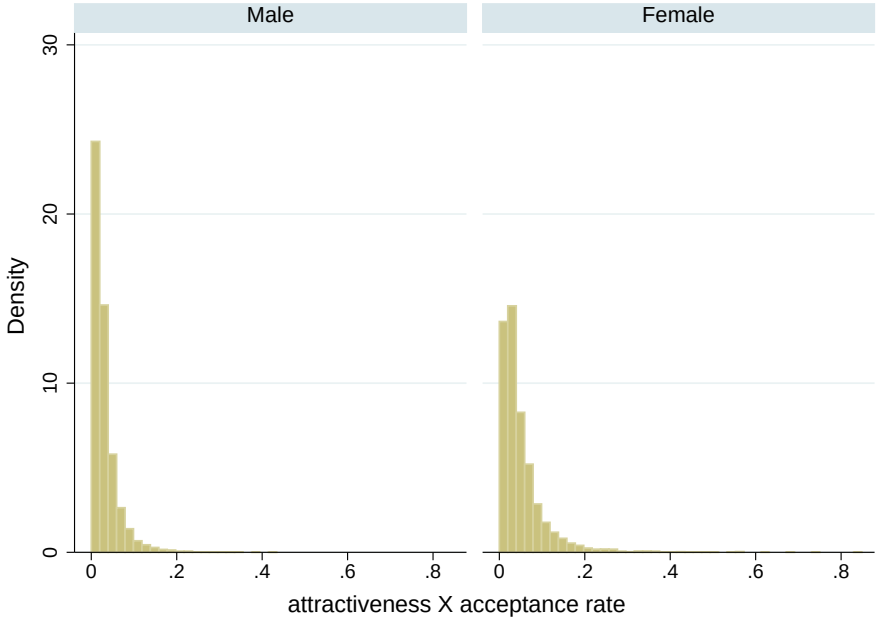
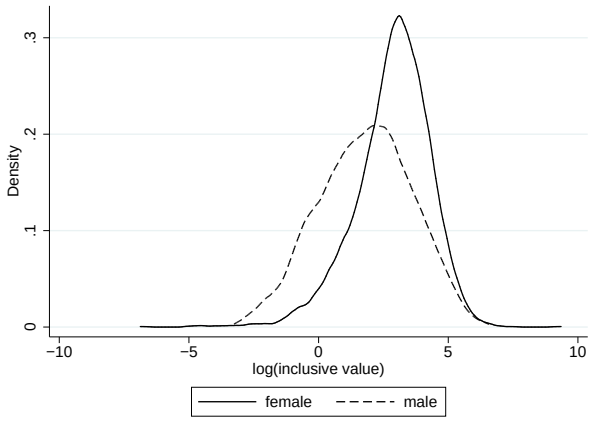
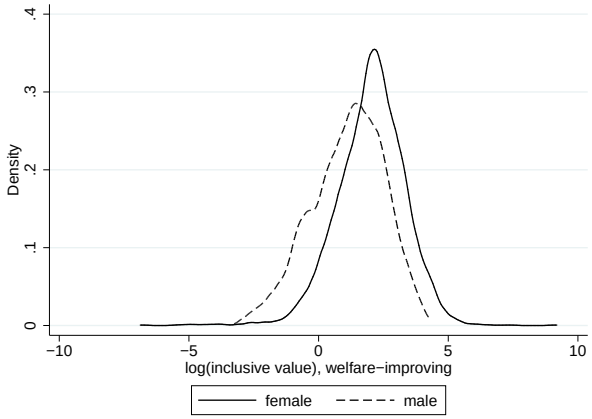


Figure 2.9: Probability of a match (attractiveness×acceptance rate), by gender

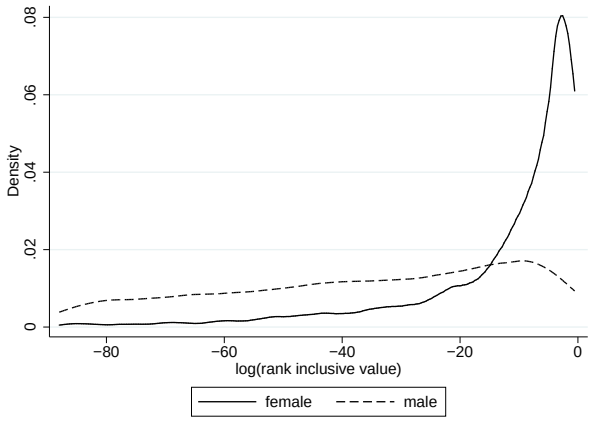


(a)

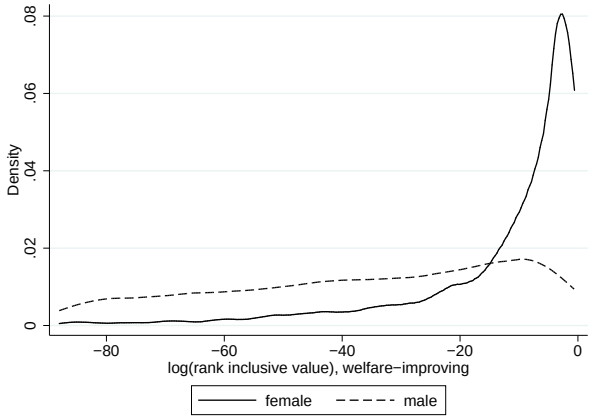


(b)

Figure 2.10: Distribution of inclusive values, by gender



(a)



(b)

Figure 2.11: Distribution of rank inclusive values by gender

Chapter 3

Information transmission in high dimensional choice problems: The value of online ratings in the restaurant market

Acknowledgements: I would like to thank Johannes Kunz, Janina Nemitz, Rainer Winkelmann, Ali Yurukoglu, the participants of the 2014 Zurich Workshop on Economics, the participants of the 2016 Annual Conference of the German Economic Association and seminar participants at the University of Zurich for helpful comments and suggestions. Errors and omissions are my own.

3.1 Introduction

Choosing a place to eat is not easy. According to Beinhocker (2007), there exist over 50'000 restaurants in New York City alone. Restaurant choice is a demanding, highdimensional choice problem from the perspective of an individual - so much so that it has been argued (Simon, 1955; Ormerod et al., 2012; Sela and Berger, 2012) that with such an abundance of choices, individuals may lack the processing capacity to select the optimal choice, even when (or indeed especially when) complete information is available. Consequently, it has been suggested that individuals may therefore resort to other strategies to make their decisions, such as heuristics. In the case of restaurant choice, uncertainty about the payoffs is further complicating the problem, as not only alternatives are abundant, but also the quality of the food is only imperfectly observable until after the choice was made.[1]

It is increasingly recognized that in the context of such problems, social interactions may play a crucial role, possibly affecting both efficiency and equity of the resulting allocation (Granovetter, 1985; Jackson, 2010; Vega-Redondo, 2007). Instead of processing the vast amount of imperfectly observed information and making independent choices, individuals may use the observed choices of others to help guide their decisions. In what they call social network markets, Potts et al. (2008) suggest that "the very act of consumer choice is governed not just by the set of incentives decribed by conventional consumer demand theory, but by the choices of others in which an individual's payoff is an explicit function of the action of others."[2] In other words, each individual's action creates positive externalities, transmitting information about an alternative from which future decisionmakers can learn (De Vany and Walls, 1996). This in turn leads to correlated choices and individuals clustering around a limited set of the many choices available. Social interactions have been shown to play a significant role in a variety of contexts, including movie atten-

[1] This is true even for repeated visits of an individual at the same restaurant, as food quality may vary from day to day or meal to meal.

[2] Note that I will use the terms "information transmission", "observation of actions of others" and "social interactions" as a synonym for others' choices.

dance (De Vany and Walls, 1996; McKenzie, 2008), music (Salganik et al., 2006), book sales (Beck, 2007), health care plan choice (Sorensen, 2006) and hospital choice (Guimaraes and Lindrooth, 2007; Pauly and Satterthwaite, 1981). Becker (1991) made a similar argument for restaurants in a theoretical framework.

In this paper, I analyze the role of information transmission via social interactions in the context of restaurant choice. More specifically, I put the focus on two aspects. First, I show that overdispersion in restaurant checkins is present in the data and that this overdispersion can be explained using a Polya Urn dynamic, incorporating previous guests choices in the decision process. I model the strength of social interactions as a function of both information exchange-related as well as socioeconomic group variables. To my knowledge, this paper is the first to embed social interactions in this manner and put the Polya Urn dynamic to use in the context of restaurant choice.

Second, I analyze the importance of user-provided ratings and other attributes of restaurants in the social interactions model. The importance of user-provided ratings has been hotly discussed in recent years both in popular media and research (for an example in the restaurant context, see Luca (2011)), with such ratings sometimes being described as a new form of currency or reputation for businesses. While debated, an assessment of the value of such ratings remains an open issue, and especially in a context with correlated behavior such ratings may provide little insights with respect to the economic outcomes of such businesses, as in these cases individuals follow the previous choices regardless of objective measures. Put simply, in a world where choices are perfectly correlated, a first individual chooses a restaurant and all successors adopt the same choice no matter what the rating says (whereas in the case of independent choices, individuals decide solely on attributes of a restaurant, disregarding the choices of others).

Researchers analyzing choice problems with aggregate data typically use McFadden's random utility framework and estimate a conditional logit model (McFadden, 1974). However, from an econometric perspective, social interactions lead to overdispersion and violate

the multinomial assumption, resulting in a misspecified model. To account for information transmission from other individuals and the resulting overdispersion, I use a Dirichlet multinomial regression model, which treats choice probabilites as random variables rather than as constant parameters of the multinomial distribution. Under the Dirichlet multinomial distribution, conditional choice probabilities increase proportionally to the number of individuals who have previously chosen that alternative, introducing a sequential aspect by making todays choices dependent on past choices - even in the case when one only has cross-sectional data at hand. The multinomial distribution and thus independence of individual choices is nested in the Dirichlet-multinomial as a limit case, which allows directly testing the two models against each other.

While the Dirichlet-multinomial model has been used in the econometric literature before (e.g. Guimaraes and Lindrooth, 2007), researchers typically omit to back out the parameters of the Dirichlet distribution, thereby failing to draw final conclusions about the predictability of outcomes. As I am also interested in the predictability of success of a restaurant dependent on its rating, I will discuss these parameters in this paper.

I use a dataset provided by the online urban guide Yelp. The dataset contains information on checkins, ratings and other attributes of restaurants in the metropolitan area of Phoenix (AZ), covering 3,171 restaurants across 125 ZIP codes within a period from early 2010 to early 2015. While it may be true that platforms such as Yelp offer only selected samples, I regard an ecosystem with an average 135 million monthly users and 67 million reviews as being interesting in itself. The fact that such platforms impact real-world decisions and their relevance is growing as the use of these platforms spreads across the population makes the study of such data important and meaningful beyond its own sake.

The model is estimated across different markets and allows the social interaction parameter to vary across these markets. By modelling social interactions within a market directly as a function of market level variables, I can broadly separate effects of within-market homogeneity of individuals from social interaction effects. For example, one would

expect that in a market with low income inequality, interactions are higher than in markets with higher income inequality, as individuals are more alike (in terms of income) and may therefore learn more from their peers than in a high-inequality environment.

In the baseline specification, markets are defined on a ZIP code and price category level (each price category in each ZIP code constitutes a market). To check for the robustness of results, different market definitions are used, markets are evaluated across and within three time periods and choices are analysed on a higher aggregation level. All the results are robust vis-à-vis these alternating definitions and periods.

I find strong evidence for social interactions, justifying the use of the Dirichlet multinomial model. The interactions are driven by a combination of factors, including both information transmission variables (such as the total number of reviews in a market) and market-characteristic variables (such as the price level, the income level or the number of competitors). Meanwhile, within-group heterogeneity in income as measured by an inequality-proxy has no significant impact on correlation within a group. I also find that higher ratings have a positive effect on visit probabilities, but that the overall informational value of such ratings is limited.

Section 2 explains the role of internet data in information transmission, section 3 discusses data, followed by the outline of the model and estimation of the model in section 4. Section 5 presents main results, section 6 shows robustness checks. Section 7 concludes.

3.2 Information transmission and the role of internet data

In the classic perspective of choice theory, individuals observe a set of choices and their characteristics before choosing whatever option maximizes their utility. The restaurant market can be seen in such a choice context as well, but it has two features that differentiate it from more traditional choice applications. For one, there is uncertainty about the choices'

arguably most important characteristic: the quality of a restaurant (including the food, the location and/or the athmosphere). A priori, individuals do not have certainty about the payoffs of eating at a particular restaurant. An individual may know that she has a preference for pizza, but she does not know whether she will like the pizza served at a particular restaurant. All she can do is form an expectation about the quality based on restaurants' observables, which might be more or less accurate. At the same time, the a priori unobserved or only imperfectly observed quality component seems crucial in explaining the highly unequal economic outcomes of restaurants (as depicted by the checkins distribution on the left hand side in Figure 3.1). This is true even within a geographic area, price or food category.

The second particularity in the restaurant context is the abundance of choice. While in theory this does not change the choice problem, it has been argued that from a behavioral perspective, the abundance of alternatives and the corresponding information attached to these options is so vast that it makes it impractical for individuals to sift through all the options before making a choice (Simon, 1955). This is not just true in contexts of incomplete or imperfect information, but also in situations where all the relevant choice attributes are observed - indeed, complete information might worsen the problem as it raises the computational burden.

Researchers have long argued that both in contexts with uncertain payoffs as well as in contexts with high-dimensional choice sets, individuals resort to social interactions (Ormerod et al., 2012), either to learn new information from signals of others or to circumvent the burden of evaluating all the information themselves[3]. The interactions can broadly be thought of as externalities: every time an individual takes a decision, she gives away information to her peers that they in turn can take into account in their own decisions. Building on the previous choices, individuals deciding later in the sequence can filter out the best alternative. Repeating this process induces correlation across the decisions of

[3]Social interactions have many labels such as information transmission, information diffusion or herding; I will use these terms interchangeably here.

individuals, which in turn can lead to highly unequal outcome distributions.

Such choices and interactions have been hard to measure and track in the past. But in recent years, internet data has been growing at an unprecedented pace. For example, by logging checkins and reviews of their users, Yelp measures and publishes information that was largely restricted to observations in limited geographic space (where people go) and word of mouth (what they tell about their experience) before the internet age. It is by using this data that I want to analyze how individuals choose restaurants, in light of the arguments made above.

The paper focuses on two measures in Yelp data: checkins and ratings. By including previous checkins of others in the decision process, one induces correlation across individuals' choices (or the same individual choosing multiple times), enabling the "rich-get-richer"-property that leads to the unequal outcome distribution seen in Figure 3.1. Ratings, on the other hand, are used as a proxy for the quality of a restaurant (assumed to be unbiased), where higher quality leads to a priori higher checkin probabilities, ceteris paribus. While these ratings are a measure for quality, they can be noisy, especially in the beginning when the overall rating is based on only a few observations: Different individuals may rate the same restaurant very differently and may therefore not rely on these ratings all that much. In an extreme case, individuals may disregard others' opinions completely, deciding in complete uncertainty with respect to the quality of a restaurant. At the other end of the limit case, ratings are taken as an objective and valid measure of quality, in which case no person should go to a low-rated restaurant as long as there is a higher-rated one (conditional on cost). Data on the choices of individuals in combination with the model described below allow to assess the informational "value" of these ratings (where a rating is of no value in the former extreme case of complete uncertainty, but very valuable in the latter case). As individual ratings accumulate over time, the average rating should become more accurate, and one should move away from the first extreme, towards the second extreme. This dynamic should be detectable when analyzing the restaurant market

at different points in time.

3.3 Data

Yelp is an online urban guide collecting visit and review data on businesses, most of which are active in the food and drink industry. The platform is visited by about 6 million people daily. The audience is characterized by an overweight of female visitors, an overweight in the 18 to 44 year old group and an above-average eduation and income level, relative to demographics of the average web user in the US.[4]

The data used in this paper is a collection of samples offered through the Yelp dataset challenge and includes information on businesses in the metropolitan area of Phoenix, Arizona, covering three snapshots over a period of five years between early 2010 and early 2015. Data on checkin counts (a feature that Yelp introduced in January 2010) as well as ratings and other restaurant attributes have been accumulated over the period from January 2010 to January 2015, with the snapshots taken in early 2013, early 2014 and early 2015. I use the checkins as a count measure for restaurant visits. Snapshots show cumulative checkins from 2010 onwards; having several snapshots across time allows me to calculate annual checkins for the years 2013 and 2014. The 3,171 restaurants in the final dataset span 42 cities, 125 ZIP code areas and count a total of 895,265 checkins as of early 2015. Roughly half of the restaurants are chain restaurants (defined as restaurants existing in more than 1 location). Additional data on ZIP-level economic and demographic characteristics are 5-year estimates from the 2013 US Census Survey.

Individuals checking in at a restaurant inform others about their choice. The probability of choosing a particular restaurant is calculated as the number of Yelp checkins divided by the total number of checkins in the same market. This checkin-based definition is different from more traditional measurements such as revenue-based market share calculations, and

[4] `https://www.quantcast.com/yelp.com`

as such it has its own characteristics. Rather than seeing checkins as a 1:1 reflection of visits, one should see it as a proxy for these visits. The number of checkins, averaging somewhere below 300 at the end of the sample period, is (presumably) much lower than the actual number of visits. Also, in my simple model, the unconditional probability of an individual checking in on Yelp is assumed to be constant across individuals, restaurants and time. This is clearly a strong simplification, and a violation of this assumption may introduce measurement error. If the measurement error is independent of the true probability p_j, it must be true that the variance of the observed probability is higher than the true variance. Alternatively, if the measurement error is like prediction error, that is, the observed choice probability is an unbiased prediction of the true probability, then the observed variance underestimates the true variance (Glaeser et al., 1996). Yelp users might differ with respect to the checkin probability among themselves as well as compared to non-Yelp users. Also, checkin probabilities might differ across restaurants, presumably overestimating true visit probabilities of trendy or especially good restaurants while underestimating true probabilities of less popular alternatives.

That being said, it is important to note that I am mainly interested in the behavior and the dynamics within the Yelp ecosystem; after all, Yelp users have access to that same information to make their decisions. In the end, these decisions happen in the "real" world, and actions based on information from online platforms become increasingly relevant as the use of such platforms spreads across the population. Alternatively, one can interpret a checkin of an individual as an implicit recommendation to others. While such a viewpoint is different, it serves equally well to a researcher interested in information transmission. Aside from that, more traditional proxies for market shares or choice probabilities do not come without their own drawbacks. Both revenue-based and profit-based calculations tend to overweigh expensive, high-margin restaurants, for example.

The restaurant population is restricted to restaurants that are recorded over the whole period of 2010 to 2015. A first look at the data in Figure 3.1 confirms the clustering

around a few choice alternatives found in other social network markets. The distribution of cumulative checkins over a period of five years shows a large number of restaurants with very few checkins, and a small number of restaurants collecting a disproportionately high number of checkins: The top 10 percent of restaurants collectively combine as many checkins as the bottom 57 percent. That inequality is not reflected in the (bounded) ratings distribution shown in the right half of Figure 3.1, which is roughly bell-shaped and centered around a mean value of 3.5. Distributions on lower aggregation levels are characterized similarly. Note that, on an individual restaurant basis, the variance of the ratings decreases the higher the average rating, an indication that individuals generally agree on what is a good restaurant, but have different opinions when it comes to bad ones (not shown here).

There's a set of 8 variables used in estimation, with a number of additional interaction variables (price category and rating, price category and income). Summary statistics for all variables are shown in Table 3.1. The dependent variable CUMULCHECKINS measures the cumulative checkins at a restaurant. The checkin distribution is highly skewed, with few restaurants capturing a disproportional share of the market (see also Table B.1 in the appendix). Note that checkins have no exact timestamp, but can be assigned to the years 2010-2012, 2013 or 2014. The main explanatory variable is the STARS rating, the (rounded) average rating given by users which is used as an intial quality guess (which supposedly gets better as the number of ratings grows). Ratings go from 1 to 5 in half-steps, are constant over time in the vast majority of cases and will serve as a measure of quality of a restaurant. As there are only few restaurants in the lower and upper parts of the STARS distribution, the variable has been recoded as a rating variable with four groups defined as $RATING1 = \{1, 1.5\}, RATING2 = \{2, 2.5\}, RATING3 = \{3, 3.5\}, RATING4 = \{4, 4.5, 5\}$. The variable PRICE indicates the price category of a restaurant, ranging from 1 to 4. Price category one is a restaurant serving food below 10 dollars, price category two ranges from 11 to 30 dollars, category three from 31 to 60 dollars and category four is for

prices from 61 dollars. The categories three and four have been merged due to the small number of restaurants in these categories. The variable LNSUMREVIEWS is defined as the logarithmized total number of reviews within a group, serving as a measure for the overall Yelp activity within the group. The variable LNCOMP is the logarithmized count of the number of competitors in the same group. The variables LNINC and LNPOP are ZIP-level measures for log-income and log-population, respectively. Finally, the variable INEQ is an income inequality measure defined as the ratio of mean and median income (ZIP-level) used to approximate within-group homogeneity. An inequality measure above 1 implies a right-skewed income distribution. Within-group homogeneity is potentially important since in a more homogenous group, individuals are expected to respond more strongly to the information provided by others, while in more heterogeneous groups such signals may not be as important.

3.4 Model

The model of restaurant choice employed here is an adapted version from De Vany and Walls (1996), who used it to explain the unequal revenue distribution at the box office via word-of-mouth recommendations and correlated decisions of moviegoers. The model is a generalized version of a Polya Urn scheme, where the first individual draws a ball of color j with some probability from an urn, replaces that ball and adds an additional ball of the same color to the urn, thereby enabling a "rich-get-richer" dynamic that is captured as social interactions in the present context. It is also closely related to the Chinese Restaurant Process (Aldous, 1985), but with a preset number of restaurants (blocks in the partition).

3.4.1 Restaurant choice with perfect information

I start with the benchmark of a simple, perfectly informed world. Suppose there is a sequence of $i = 1, \ldots, N$ individuals in a market M who have to choose a place to eat

from a choice set of R restaurants. The decision set of the ith consumer is denoted by $d_i = \{1, \ldots, R\}$ and is identical for all individuals in the sequence. For simplicity, assume that each of these restaurants represents a distinct quality level q. Individuals maximize a utility with a quality-dependent payoff.

In the case of perfect information, all individuals make choices observing quality perfectly, leading to a quality-dependent vector of choice probabilities $p(q) = \{p_1, \ldots, p_R\}$ that is equal across individuals. The N individuals make their choices independently from one another and allocate themselves across restaurants $1, \ldots, R$, resulting in the allocation vector $A = \{A_1, \ldots, A_R\}$ whose realization can be defined as $a = \{a_1, \ldots, a_R\}$ where

$$\sum_{k=1}^{R} a_k = N.$$

Any particular outcome is multinomial distributed with probability function

$$\begin{aligned} &Pr\{A_1 = a_1, \ldots, A_R = a_R | p_1, \ldots, p_R\} = \\ &Pr\{A = a|p\} = \frac{N!}{a_1! \ldots a_R!} p_1^{a_1} \ldots p_R^{a_R} \end{aligned} \tag{3.1}$$

3.4.2 Restaurant choice with imperfect information

Now suppose that the quality of a restaurant is unknown or only imperfectly known to individuals at the moment they make their decisions, introducing uncertainty about the payoffs of a particular choice. Instead of maximizing utilities, individuals now maximize expected utilities that, as I will show later, depend on their position in the choice sequence.

While in this setting, individuals cannot observe the quality of a restaurant directly, they now have two alternative measures at their disposal. For one, they can form an expectation about the quality of a restaurant using restaurant ratings $\mathrm{E}(s) = q$, which on average is assumed to be an unbiased quality estimate. On the other hand, individuals can observe the choices of the $(i-1)$ individuals preceding them in the sequence and use these observations as additional signals for quality.

In statistical terms, p is now itself a random vector dependent on ratings, rather than a constant parameter reflecting quality as in the perfect information case. Whereas before, a was the random realization of an allocation vector generated by the parameters in p, there is now an additional random layer as the entries in p are generated from a random process as well.

The probabilty density of choosing a restaurant given p, $P(A = a|p)$, is still a known function: It's the multinomial density shown in the previous section. To account for the fact that p is a random vector, I need a prior probability distribution for p supposed to reflect individuals' beliefs about the ratings as a quality indicator. In other words, as p is a function of ratings, one can interpret this prior probability distribution as the informational value of a rating, absent any information about the choices of others.

Now consider what happens when individuals later in the sequence start taking into account the choices that were made previously. They can use the previous choices to update their beliefs and form a posterior distribution, where the posterior distribution of p is proportional to the product of the known density of choosing a restaurant given p and the prior distribution of p. One can now see what role previous choices play: They change the distribution of the number of guests visiting a restaurant. For example, it can be shown that on average, the distributions become more concentrated, reflecting an average gain in knowledge (Lindley, 1961). So while the first individual in a sequence can base his decision only on a prior distribution, the individual making the Nth decision has access to much more accurate information to make his decision.

$P(A = a|p)$ is still the multinomial density defined previously. What I need to embed quality uncertainty is a prior distribution of p given the ratings, $P(p|\delta^{-1}\alpha)$, where the value of a rating is captured in the parameters δ and α. In my case, I assume p to be Dirichlet distributed. The Dirichlet distribution (also known as the multivariate Beta distribution) is a conjugate prior for the probability parameter p of the multinomial distribution with

density:

$$f(p_1, \ldots, p_R; \alpha_1, \ldots, \alpha_R, \delta) = \frac{1}{B(\delta, \alpha)} \prod_{k=1}^{R} p_k^{\delta^{-1}\alpha_k - 1} \tag{3.2}$$

where $\sum_{k=1}^{M} p_k = 1, \alpha_k > 0, \delta > 0$. $B(\delta^{-1}\alpha)$ is the β function that can be expressed in terms of the Gamma function:

$$B(\delta, \alpha) = \frac{\prod_{k=1}^{R} \Gamma(\delta^{-1}\alpha_k)}{\Gamma(\sum_{k=1}^{R} \delta^{-1}\alpha_k)} \tag{3.3}$$

The marginal Beta distribution of the Dirichlet distribution (Beta$(\alpha_k, \sum_{k=1}^{R} \alpha_k - \alpha_k)$) has mean

$$\mathrm{E}(p_j) = \frac{\delta^{-1}\alpha_j}{\sum_{k=1}^{R} \delta^{-1}\alpha_k} = \frac{\alpha_j}{\sum_{k=1}^{R} \alpha_k}. \tag{3.4}$$

The covariance between the choice probabilities for restaurants j and m is given by

$$Cov(p_j, p_m) = \frac{-\alpha_j \alpha_m}{(\sum_{k=1}^{R} \alpha_k)^2 (\sum_{k=1}^{R} \delta^{-1}\alpha_k + 1)}, \quad j \neq m \tag{3.5}$$

When $R = 2$, the Dirichlet reduces to the Beta distribution.

Figure 3.2 shows distributions over p_j for different values of δ (for simplicity, $\alpha_k = \alpha = 1$, $\mathrm{E}(p_j) = 0.5$). When all components of $\delta^{-1}\alpha$ are equal to 1, the Dirichlet distribution reduces to the uniform distribution over the probability simplex. When the components of $\delta^{-1}\alpha$ are all greater than 1, the density is unimodal, and when the components of $\delta^{-1}\alpha$ are all less than 1, the density has sharp peaks at the boundaries. As $\delta \to 0$, the distribution over p degenerates to a constant, i.e. there is no uncertainty about the quality left and we're back in the perfect information case with the multinomial distribution.

Up to now I only considered a single market. To uncover parameters of the distribution, I will need to estimate across multiple markets, where restaurants with identical ratings are assumed to be identical in terms of expected quality (both within as well as across markets). This may sound simplistic and reductionist, but actually reproduces the choice

problem faced by individuals: Before actually having eaten at a restaurant, a rating and some rudimentary information such as a price range is all an individual can base his decision on.

Consider a set of markets $m = 1, 2, \ldots, M$. For the mth market, there is an associated vector $p^{(m)}$ of length $k = R$ for the probabilities of going to each restaurant. Suppose that one can model these M probability vectors as coming from a $\text{Dir}(\delta^{-1}\alpha)$ distribution and that we have N_m samples from the mth probability vector. This prior distribution is then compounded with the multinomial distribution. The resulting $\{a_m\}$ are realizations of a Dirichlet-multinomial distribution.

The $m = 1, 2, \ldots, M$ sets of samples $\{a_m\}$ drawn from the M probability mass functions drawn from the $\text{Dir}(\delta^{-1}\alpha)$ distribution are conditionally independent given $\delta^{-1}\alpha$, so the likelihood of $\delta^{-1}\alpha$ can be written as the product

$$Pr\{A = a\} = \prod_{m=1}^{M} Pr\{A_m = a_m|\delta^{-1}\alpha\} \tag{3.6}$$

where the compounded, unconditional distribution of A_m is obtained by integrating over p:

$$\begin{aligned} Pr\{A_m = a_m|\alpha\} &= \int Pr\{A_m = a_m|p\} f(p|\alpha) dp \\ &= \frac{N_m!}{\prod_{k=1}^{R} a_{mk}!} \frac{\Gamma(\sum_{k=1}^{R} \delta_m^{-1}\alpha_k)}{\Gamma(N_m + \sum_{k=1}^{R} \delta_m^{-1}\alpha_k)} \prod_{k=1}^{R} \frac{\Gamma(a_{mk} + \delta_m^{-1}\alpha_k)}{\Gamma(\delta_m^{-1}\alpha_k) a_{mk}!} \\ &= \frac{NB(\delta^{-1}A_m, N_m)}{\prod_{k:a_{mk}>0} a_k B(\delta_m^{-1}\alpha_k, a_{mk})} \end{aligned} \tag{3.7}$$

where $NB()$ denotes the negative binomial distribution. By substituting Equation (3.7) into Equation (3.6), one obtains the likelihood of the observed data (see Equation 3.12 in the estimation section).

Using the law of iterated expectations and the law of total variance, the Dirichlet multinomial can be shown to have expected value, variance and covariance given by (Ng

et al., 2011)

$$\mathrm{E}(a_{jm}) = \mathrm{E}[\mathrm{E}(a_{jm}|p_{jm})] = N_m \mathrm{E}(p_{jm}) = \frac{N_m \alpha_j}{\sum_{k=1}^{R} \alpha_k}, \tag{3.8}$$

$$\begin{aligned} Var(a_{jm}) &= Var(\mathrm{E}[a_{jm}|p_{jm}]) + \mathrm{E}[Var(a_{jm}|p_{jm})] \\ &= \mathrm{E}(N_m p_{jm}(1 - p_{jm})) + Var(N_m p_{jm}) \\ &= N_m \frac{\alpha_{jm}}{\sum_{k=1}^{R} \alpha_k} \left(1 - \frac{\alpha_j}{\sum_{k=1}^{R} \alpha_k}\right) (1 + (N_m - 1)\rho_m) \end{aligned} \tag{3.9}$$

where $\rho = 1/(\delta_m^{-1} \sum_{k=1}^{R} \alpha_k + 1)$ is an overdispersion parameter, inflating the variance by a factor $(1+(N_m-1)\rho_m)$ vis-à-vis the multinomial distribution. Note that α_j is restaurant-specific, while δ_m is market specific. The expected value of the Dirichlet-multinomial distribution is independent of δ; in the case of the variance, δ enters through ρ. In the case of perfect information and no uncertainty about payoffs, $\delta \to 0$, $\rho \to 1$ and the overdispersion disappears.

It is important to point out that the social interactions as modeled here could be interpreted both as a form of information transmission or individuals simply having a preference to be surrounded by other individuals (or a combination of the two). In the former case individuals indirectly find out about the quality of a restaurant, while in the latter case the presence of others itself becomes an attractive feature of a restaurant. When only looking at choices on the level of individual restaurants, both explanations are observationally equivalent in the sense of Ellison and Glaeser (1997).

3.4.3 Parameters of interest

To see the role previous choices play in this process, it is instructive to look at the conditional probability of the $(N+1)$st individual to eat at restaurant j given the first N individuals have led to the allocation vector $A_N = a_N$. For notational ease, I drop the m-subscript and go back to focussing only on a single market. If $(A|p) \sim \text{Multinomial}_k(N, p)$

and $P \sim \text{Dir}(\delta^{-1}\alpha)$, then $(P|A = a) \sim \text{Dir}(\delta^{-1}\alpha + a)$. Based on this, the conditional probability given the first N individuals can be shown to be equal to

$$\begin{aligned} Pr\{d_{N+1} = j|A = a\} &= \frac{\delta^{-1}\alpha_j + a_j}{N + \sum_{k=1}^{R} \delta^{-1}\alpha_k} \\ &= \frac{\delta^{-1}\alpha_j}{\sum_{k=1}^{R} \delta^{-1}\alpha_k} \frac{\sum_{k=1}^{R} \delta^{-1}\alpha_k}{N + \sum_{k=1}^{R} \delta^{-1}\alpha_k} + \frac{a_j}{N} \frac{N}{N + \sum_{k=1}^{R} \delta^{-1}\alpha_k} \\ &= w\text{E}(p_j) + (1 - w)\frac{a_j}{N} \end{aligned} \tag{3.10}$$

where $w = \frac{\sum_{k=1}^{R} \delta^{-1}\alpha_k}{N + \sum_{k=1}^{R} \delta^{-1}\alpha_k}$. The last step in Equation (3.10) decomposes the probability into a weighted average of a prior probability and a likelihood component. The first line in Equation (3.10) nicely illustrates how information transmission depends on α and δ. The parameter vector α serves as a (conditional) quality measure (higher quality restaurants have higher α_j), which is scaled by δ on a market level. The conditional probability also illustrates an important property of the Dirichlet-multinomial distribution: While the individuals' decisions are (clearly) not independent, they are exchangeable, meaning that the order in which individuals choose is irrelevant.

A higher α_j increases the expected probability of an individual visiting the restaurant relative to restaurants with a lower rating, ceteris paribus. If social interactions are strong enough though, these relative differences may become irrelevant. This is reflected in the δ parameter. δ does not influence the marginal expectation of any alternative, but does influence the marginal variance, capturing social interactions. As $\delta \to 0$ ($\delta^{-1} \sum_{k=1}^{R} \alpha_k \to \infty$), the Dirichlet multinomial converges to the multinomial distribution with a constant p vector and independent choices. Differences in choice probabilities are entirely determined by differences in ratings refleted in α_j - reliance on social interactions becomes irrelevant.

$\delta \to \infty$, on the other hand, leads to a sparse distribution of p and a process heavily determined by social interactions (i.e. almost all people go or don't go to a particular restaurant, irrespective of quality or other attributes). In other words, the first individual flips a coin and chooses a restaurant at random, while all others following in the sequence

adopt the first individual's choice.[5] Outcomes become highly skewed, as they are heavily influenced by strong but a priori unpredictable social interactions. The α vector completely looses its predictive power. A special case is $\delta^{-1}\alpha_j = 1 \; \forall \; j$ with a uniform distribution over outcomes, i.e. all possible outcome allocations are a priori equally likely (also known as the Bose-Einstein distribution, discussed in De Vany and Walls (1996)).

3.4.4 Grouping restaurants by quality

Up to this point, the model assumed that each of the R restaurants represents a distinct quality level. When estimating the model, this will not be the case, as quality will be measured by a rating with a 9-point scale only. I will estimate the model on the individual restaurant level, assigning the same α_j parameter to restaurants with the same rating, making restaurants with the same rating are a priori indistinguishable from each other. This may seem like an oversimplification, but for the ratings to be of any value a high rating in one place should translate into an approximately equivalent signal for quality in another place (conditional on location and price category). If ratings and their informational value are idiosyncratic, they are of little help.

Alternatively, I also aggregate restaurants by their quality and estimate the model, taking these aggregated restaurant groups as the observational unit. With respect to the Dirichlet-multinomial distribution, partitioning $\{1, 2, \ldots, R\}$ restaurants into $\{C_1, C_2, \ldots, C_s\}$ with $s < R$ is straightforward, as

$$(\sum_{i \in C_1} P_i, \sum_{i \in C_2} P_i, , \ldots, \sum_{i \in C_s} P_i,) \sim \text{Dir}(\sum_{i \in C_1} \alpha_i, \sum_{i \in C_2} \alpha_i, \ldots, \sum_{i \in C_s} \alpha_i)$$

due to the Dirichlet's aggregation property (Frigyik et al., 2010). By definition then, the α_j in the aggregated model should be at least as high as the in the model working with

[5] In such a case, the model would degenerate to a constant-only conditional logit model with restaurant fixed effects - which could be interesting from a marketing perspective (e.g. is the constant for restaurant A higher than the constant for restaurant B), but not interesting in the context of social interactions where individuals only have limited access to information. Also, overdispersion would be ignored.

individual restaurants.

3.4.5 Estimation

The model derived in the previous section can be estimated by maximum likelihood (Guimaraes and Lindrooth, 2007). Main results will focus on markets defined on a ZIP code and pricerange level, thereby assuming information is exchanged within a given price range and limited geographic area (i.e. social interactions are limited within that group). Since I do not observe any individual-specific information, actual restaurant choices can be aggregated into a vector of counts $a_m = \{a_{1m}, \ldots, a_{Rm}\}$ without any loss of information. Consequently, individuals in a given market face the same choice set with identical choice attributes. As noted previously, individual visits are not independent but exchangeable, making the counts independent of the ordering of individuals in the queue while maintaining the sequential nature of the choice problem.

As shown by Guimaraes and Lindrooth (2007), modelling p_{jm} as a random variable is equivalent to introducing unobservable market-specific effects that equally influence the decisions of all individuals belonging to the same market. These market-specific effects will induce correlation across individuals in the same market, which in turn leads to overdispersion in the a_{km} count. The choice probability of individual i in market m selecting choice j (conditional on the group random effects) is defined as

$$p_{ijm} = \frac{\exp(\beta' \mathbf{x_j} + \eta_{jm})}{\sum_{k=1}^{R} \exp(\beta' \mathbf{x_k} + \eta_{km})} = \frac{\alpha_j \exp(\eta_{jm})}{\sum_{k=1}^{R} \alpha_k \exp(\eta_{km})} \tag{3.11}$$

where $\alpha_j = \exp(\beta' \mathbf{x_j})$, $\mathbf{x_j}$ are observable characteristics of choice j, η_{jm} are the random effects that affect members in market m and ε_{ijm} are assumed to be independent conditional on the market random effects. Assume that the random market effects $\exp(\eta_{jm})$ are i.i.d. gamma distributed with parameters $\{\delta_m^{-1}\alpha_j, \delta_m^{-1}\alpha_j\}$ where $\delta_m > 0$ is a market-specific parameter. Then, $\exp(\eta_{jm})$ has unit expectation and a variance equal to $\delta_m \alpha_j^{-1}$. Moreover, the variables defined by the product $\alpha_j \exp(\eta_{jm})$ also follow independent gamma

distributions with parameters $\{\delta_m^{-1}\alpha_j, \delta_m^{-1}\}$. Given that all these variables follow independent gamma distributions with the same scale parameter, the vector $p = \{p_{1m}, \ldots, p_{Rm}\}$ follows a Dirichlet distribution with a density as defined in Equation 3.2 (Mosimann, 1962; Guimaraes and Lindrooth, 2007).

Compounding the Dirichlet with the multinomial and involving all markets M leads to the unconditional likelihood function

$$\begin{aligned} L_{DM} &= \prod_{m=1}^{G} \int \prod_{k=1}^{R} a_m! \frac{p_{km}^{a_k m}}{a_{km}!} f_{DM}(p_{1m}, \ldots, p_{R-1m}) dp_1 \ldots dp_{R-1} \\ &= \prod_{m=1}^{G} \frac{a_m! \Gamma(\delta_m^{-1} \sum_{k=1}^{R_m} \alpha_k)}{\Gamma(\delta_m^{-1} \sum_{k=1}^{R_m} \alpha_k + a_m)} \prod_{k=1}^{R} \frac{\Gamma(\delta_m^{-1}\alpha_k + a_{km})}{\Gamma(\delta_m^{-1}\alpha_k) a_{km}!} \end{aligned} \tag{3.12}$$

where $p_R = 1 - \sum_{k=1}^{R-1} p_k$. If the market random effects η_{jm} have a variance of zero and the correlation coefficient tends to zero, the likelihood function of the Dirichlet multinomial collapses into the likelihood function of the multinomial logit model (or grouped conditional logit model). Testing for the existence of social interactions (i.e. testing for $\delta_m > 0$) can therefore be implemented directly via a likelihood ratio test. Note that the null hypothesis for the test is in the boundary of the parameter space, and therefore the correct p-value is one-half that which is obtained from the χ_1^2 (Self and Liang, 1987; Gutierrez et al., 2001).

Guimaraes and Lindrooth (2007) show that the likelihood in Equation 3.12 can be reformulated as the fixed effects negative binomial model developed by Hausman et al. (1984), which makes estimation of the Dirichlet multinomial model readily implementable in standard statistical software packages (see also Guimaraes, 2005). As shown by Guimaraes and Lindrooth (2007), Guimaraes (2005) and - in the context of fixed effects negative binomial models - by Allison and Waterman (2002), the group parameter δ_m can be modelled as a function of group-level variables denoted by $\mathbf{w_m}$, i.e. $\delta_m = f(\gamma' \mathbf{w_m})$, which allows to identify the driving forces of the social interactions and the overdispersion parameter. This is in contrast to the multinomial logit model, where market fixed effects cancel out. In my case, I assume $-\ln(\delta_m) = \gamma' \mathbf{w_m}$. If those variables are restricted to a single constant, it is

implicitly assumed that all markets share a common δ ($\delta_m = \delta$).[6]

3.5 Main results

This section is structured as follows: First, I present coefficient estimates for the different rating categories in three different models. The first model is the traditional multinomial logit, where the probability vector p is treated as a constant. This model serves as a benchmark. In the second model, I estimate the Dirichlet multinomial model enforcing $\delta_g = \delta$ across all markets. The last model relaxes this assumption and models δ_g as a function of market-level variables, which allows insights into what drives the correlation across individuals. More specifically, I assume $-\ln(\delta_g) = \gamma' w_g$. Both Dirichlet-multinomial models can be tested directly using a likelihood ratio test comparing the likelihood of the Dirichlet multinomial to the likelihood of the multinomial logit which imposes $\delta = 0$ (the resulting p-value should be halfed, as outlined previously). Robustness checks using only single-period checkins rather than cumulative checkins, specifications aggregating restaurants by rating category as well as results using alternative market definitions can be found in the appendix.

In a second step, I back out the parameters of the Dirichlet multinomial model using the estimates of the model with constant δ. The parameters $\delta^{-1}\alpha$ can be interpreted in the context of the prior distribution, where the distribution over p converges to a constant as $\delta^{-1}\alpha \to \infty$, consequently indicating a high informational value for the Yelp ratings. Low parameter values, on the other hand, indicate a stronger role of social interactions.

Main results are summarized in Table 3.2, where the multinomial logit model and two Dirichlet multinomial models are estimated using cumulative checkins at three different

[6]Note that I deliberately abstain from using restaurant fixed effects in estimation. Using fixed effects would imply effects that are only unobserved by the researcher but not the individual, whereas in my case I explicitly want to allow for inidividuals making decisions in a setting of imperfect information. Also, if the user-provided ratings are informative enough, unobserved restaurant fixed effects should be reflected in those ratings (whether they are is one of the questions this paper addresses). Other unobserved factors such as characteristics of the neighborhood do not affect results as restaurants are grouped by location and price category and all restaurants in a group would be equally affected by such factors.

points in time. Main results focus on markets defined by both ZIP code and price category, i.e. each market is a unique ZIP x Price combination, resulting in 236 markets (robustness checks use different market definitions). Results for different time periods are estimated separately; note that since I use cumulative checkins for the main results, observations for later periods include checkins from all previous periods.

Coefficients on the rating variable are all positive and increasing, which is expected. Coefficients in the grouped conditional logit model are all highly significant at all points in time, whereas in the Dirichlet model the coefficient on the second rating category is not statistically significant at any conventional significance level. Put differently: Using the Dirichlet model, a higher rating does not correlate with higher choice probabilities if this rating is in the bottom half of the scale. Standard errors increase roughly by a factor of 4 or 5 compared to the multinomial logit model, depending on the period considered.

Focussing on the *Dir II* model, one can see what drives within-market social interactions (note that $\delta_g = \exp(-\gamma' w_g)$). Higher price categories, more competitors and higher income lead to a relatively stronger role for social interactions, while the total number of reviews in a market weakens them - which could be explained by ratings becoming more valuable as they are based on a larger number of opinions. Coefficients on the income inequality proxy and population are not significant. It is noteworthy that when including market level variables, the coefficient on the constant is close to zero and statistically insignificant. This opposed to the *Dir I* model where the coefficient on the constant is highly significant, suggesting that market level variables capture a large part of the social interaction parameter δ.

Comparison of the models using loglikelihood values, Pearson chi-square statistics and AIC statistics all clearly favor the Dirichlet models (note that Pearson statistics have been adjusted accordingly for the Dirichlet-multinomial; see Guimaraes and Lindrooth (2007); Mosimann (1962)). Interestingly, all these statistics change little over time in the case of the Dirichlet multinomial model, while they get progressively worse in the multinomial logit model as checkins accumulate over longer periods of time.

These statistics lead to the conclusion that social interactions (or previous checkins) do play a role in the decision process of individuals. The follow-up question then is: how much so? In order to answer this question, I back out the parameter vectors α and δ of the *Dir I* model and report them in Table 3.3. I also report the overdispersion parameter, ρ. The top half of the table reports parameters for the cumulative checkins, the bottom half for single period checkins.

The Dirichlet's $\alpha = \exp(x'\beta)$ parameters are a unidimensional function of a rating dummy, where higher β_j leads to higher α_j which in turn leads to a higher expected value of p_j, the probability of choosing a restaurant with rating j in a market. The α vector is then scaled by δ^{-1}, where $delta = \exp(-\gamma_0)$ and γ_0 refers to the constant, defining the Dirichlet parameters.[7] The existence of social interactions can be tested directly by testing the null hypothesis $\delta = 0$ or performing a likelihood ratio test, where the multinomial logit model serves as the restricted model. As already apparent in the statistics reported in the footer of Table 3.2, the hypothesis of $\delta = 0$ is clearly rejected.

As outlined in the model section, a Dirichlet parameter of 1 results in a uniform distribution of p (the Bose-Einstein distribution). Paramaters above 1 result in unimodel peaks at the expected value of p, while parameters below 1 lead to sparse distributions with peaks at the edges. Parameters are reported in Table 3.3. Figure 3.3 plots the densities of p for the different rating categories listed in Table 3.3.

As one can see from both the table and the figure, the Dirichlet parameter for highest rating category is the only one above 1 in the case of cumulative checkins. All other parameters are below one, indicating that little can be learned about the number of checkins from those ratings as checkins are mainly governed by a priori unpredictable social interactions. In the case of cumulative checkins, the combinations $\delta^{-1}\alpha$ stay roughly constant, even though both components increase over periods.

Single period checkins for the years 2013 and 2014 are displayed in the bottom half

[7] $delta = \exp(-\gamma' w_g)$ in the *Dir II* model.

of the table and are based on the result in Table 3.5 in the appendix. 2013 results use 1-year lags, while 2014 results use 2-year lags. Generally, all parameters are lower than in the cumulative case, indicating that focussing only on more recent checkins and discarding checkins in the more distant past, the importance of social interactions only rises relative to the importance of ratings.

What one can learn from the preceding table and figures is that while the quality reflected in the ratings do impact expected choice probabilities, the final allocation of checkins is far from certain. They give better-quality restaurants a headstart, but the choice probabilites within a rating category and their ultimate outcomes still vary wildly as individuals are responding strongly to signals of others. Intuitively, this social dynamic is best explained by the mechanism of Polya's urn: Each time an individual chooses a restaurant from the pool of possible candidates, the choice probability of that restaurant increases for the individuals choosing later in the sequence. As time passes, conditional choice probabilities (as defined in Equation 3.10) are less governed by the $\delta^{-1}\alpha_j$ parameters, and more by the previous choices of others.

3.6 Robustness checks

I perform a number of robustness checks. First, by the aggregation property of the Dirichlet distribution, restaurants can be aggregated and assigned to partitions (as mentioned in the Model section). For the results in Table 3.4, restaurants within a given ZIP code and price category are grouped by their rating, resulting in one count per rating in a group. By the properties of the Dirichlet distribution, aggregating should increase $\delta_g^{-1}\alpha_j$ and - as the number of choices within a group is decreased - mechanically decrease correlation across individuals, which is proportional to the number of choice alternatives. This is shown in Table 3.4. Results stay qualitatively the same, and even though you observe an increase in the parameters, social interactions still play a large role.

Second, to address potential concerns of endogeneity between ratings and visits, Table 3.5 shows results where only checkins within a given year are used in combination with lagged ratings.[8] Single-period visits have been obtained by differencing the cumulative checkins (which is why the 2010-2012 is not covered). The 2013 specification uses $(t-1)$ lags, while the 2014 specification uses both $(t-1)$ and $(t-2)$ lags. Results stay qualitatively the same as the previously reported results. The exception is the coefficient for the second rating category, which becomes even weaker as before. Again, correlation across choices is highly significant, strongly favoring the Dirichlet multinomial to the multinomial logit model. Also note that coefficients increase from 2013 to 2014 when using 1-period lags, while staying constant when comparing 2013 results to 2014 results using $(t-2)$ lags. Again, loglikelihood, Pearson and AIC statistics for the Dirichlet models show little change across periods, while they worsen quite significantly in the case of the multinomial logit.

As a last robustness check, I estimate the model under a wider grouping structure. Grouping restaurants by ZIP code and price category assumes away wider spread social interactions that might exist, either across ZIP codes or across price categories. Table 3.6 in the appendix therefore presents results of the three models when the grouping is changed to ZIP code (i.e. across different price categories), to a citywide group and to a group defined on the city and price category level (instead of the ZIP and price category level). The latter leads to little changes compared to previous estimations as shown in the middle section of the table, whereas extending the groups to multiple price categories results in lower correlation within group, which is expected. Since price categories now change within group, price variables as well as interaction variables between price category and ratings and price category and group-level income are included. Interestingly, price categories have no significant effect in the Dirichlet models.

[8]Note that ratings themselves change only little over time, and if they do mostly in the early phase when tha rating is based only on few user feedbacks.

3.7 Discussion

The aim of this paper is to explore the role of information transmission through social interactions in the choice process of individuals in a highly competitive market with a wide range of hard to distinguish alternatives, exemplified by the restaurant market. The presence of social interactions leads to correlation across individuals, highly skewed allocations and generally more uncertainty in the prediction of economic outcomes. Here, social interactions are embedded in a simple but insightful Dirichlet multinomial framework. The choice of the model turns out to be superior to the traditional multinomial logit approach based on a number of different measures.

Social interactions are found to be present in Yelp's restaurant data, and can be modelled largely as a function of variables relating to aggregate information exchange variables, while socioeconomic variables such as income or income heterogeneity within a market play a negligible role as a driver of correlation. Results are robust against both restricting choices to a single period and lagged ratings as well as to widening the market definition and aggregating restaurants into their rating levels. I also find that user-provided ratings only have a limited impact on individuals' decisions, as interpreted from the magnitude of the parameters of the Dirichlet-multinomial distribution. Generally, ratings at the top end of the scale are more informative than ratings at the bottom of the scale.

This paper only deals with counts aggregated on a restaurant level and with no information on the level of the individuals themselves. It is therefore important to note that I cannot provide details on the exact definition of the social interactions. Specifically, the social interactions found here could both be a form of information transmission or individuals simply having a preference to be surrounded by other individuals (or a combination of the two). In the former case individuals indirectly find out about the quality of a restaurant, while in the latter case the presence of others itself becomes an attractive feature of a restaurant. When only looking at choices on the level of individual restaurants, both

explanations are observationally equivalent in the sense of Ellison and Glaeser (1997).

References

Aldous, D. J. (1985). *Exchangeability and related topics*. Springer.

Allison, P. D. and Waterman, R. P. (2002). Fixed effects negative binomial regression models. *Sociological Methodology*, 32(1):247–265.

Beck, J. (2007). The sales effect of word of mouth: a model for creative goods and estimates for novels. *Journal of Cultural Economics*, 31(1):5–23.

Becker, G. S. (1991). A note on restaurant pricing and other examples of social influences on price. *Journal of Political Economy*, 99(5):1109.

Beinhocker, E. D. (2007). *The origin of wealth: Evolution, complexity, and the radical remaking of economics*. Random House.

De Vany, A. and Walls, W. D. (1996). Bose-einstein dynamics and adaptive contracting in the motion picture industry. *The Economic Journal*, pages 1493–1514.

Ellison, G. and Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *The Journal of Political Economy*, 105(5):889–927.

Frigyik, B. A., Kapila, A., and Gupta, M. R. (2010). Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and social interactions*. *The Quarterly Journal of Economics*, 111(2):507–548.

Granovetter, M. (1985). Economic action and social structure: the problem of embeddedness. *American Journal of Sociology*, pages 481–510.

Guimaraes, P. (2005). A simple approach to fit the beta-binomial model. *Stata Journal*, 5(3):385–394.

Guimaraes, P. and Lindrooth, R. C. (2007). Controlling for overdispersion in grouped conditional logit models: A computationally simple application of dirichlet-multinomial regression. *The Econometrics Journal*, 10(2):439–452.

Gutierrez, R. G., Carter, S., and Drukker, D. M. (2001). On boundary-value likelihood-ratio tests. *Stata Technical Bulletin*, 10(60).

Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r & d relationship. *Econometrica: Journal of the Econometric Society*, pages 909–938.

Jackson, M. O. (2010). *Social and Economic Networks*. Princeton University Press.

Lindley, D. V. (1961). Dynamic programming and decision theory. *Applied Statistics*, pages 39–51.

Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp. com. *Com (September 16, 2011). Harvard Business School NOM Unit Working Paper*, (12-016).

McFadden, D. (1974). Conditional logit analysis of qualitative choices. *Zarembka. P.(eds): Frontiers*.

McKenzie, J. (2008). Bayesian information transmission and stable distributions: Motion picture revenues at the australian box office*. *Economic Record*, 84(266):338–353.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, pages 65–82.

Ng, K. W., Tian, G.-L., and Tang, M.-L. (2011). *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons.

Ormerod, P., Tarbush, B., and Bentley, R. A. (2012). Social network markets: the influence of network structure when consumers face decisions over many similar choices. *arXiv preprint arXiv:1210.1646*.

Pauly, M. V. and Satterthwaite, M. A. (1981). The pricing of primary care physicians services: a test of the role of consumer information. *The Bell Journal of Economics*, pages 488–506.

Potts, J., Cunningham, S., Hartley, J., and Ormerod, P. (2008). Social network markets: a new definition of the creative industries. *Journal of Cultural Economics*, 32(3):167–185.

Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Sela, A. and Berger, J. (2012). Decision quicksand: how trivial choices suck us in. *Journal of Consumer Research*, 39(2):360–370.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, pages 99–118.

Sorensen, A. T. (2006). Social learning and health plan choice. *The Rand Journal of Economics*, 37(4):929–945.

Vega-Redondo, F. (2007). *Complex social networks*. Number 44. Cambridge University Press.

Table 3.1: Summary Statistics

	2010-2012		*2013*		*2014*	
	Mean	SD	Mean	SD	Mean	SD
CUMULCHECKINS	125.06	(218.77)	210.05	(363.23)	282.42	(480.47)
CHECKINS (1 period)	125.06	(218.77)	84.87	(154.08)	72.41	(128.49)
RATING1	0.01	(0.11)	0.01	(0.11)	0.01	(0.11)
RATING2	0.10	(0.30)	0.10	(0.29)	0.10	(0.30)
RATING3	0.47	(0.50)	0.50	(0.50)	0.51	(0.50)
RATING4	0.41	(0.49)	0.39	(0.49)	0.38	(0.49)
PRICE	1.56	(0.59)	1.56	(0.59)	1.56	(0.59)
LNSUMREVIEWS	6.51	(1.51)	6.84	(1.47)	7.11	(1.44)
LNSUMREVIEWS, adj.	3.55	(0.78)	3.88	(0.74)	4.15	(0.72)
LNCOMP	2.96	(0.95)	2.96	(0.95)	2.96	(0.95)
INEQ	1.36	(0.21)	1.36	(0.21)	1.36	(0.21)
LNINC	11.14	(0.37)	11.14	(0.37)	11.14	(0.37)
LNPOP	10.34	(0.68)	10.34	(0.68)	10.34	(0.68)
Observations	3,159		3,169		3,170	

Standard errors in parentheses. *Price*, *lncompetitors*, *lninc* and *lnpop* are constant over time and within ZIP.

Lnsumreviews and *lnsumreviewsadj* are calculated on the Price x ZIP level.

Lnsumreviewsadj is the sum of reviews in the group divided by the number of restaurants in the group.

Source : Yelp, US Census. Own calculations.

Table 3.2: Cumulative checkins over time

	2010-2012			2013			2014		
Variable	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir I*	*Dir II*
RATING2	0.537	0.178	0.248	0.866	0.277	0.333	0.634	0.229	0.309
	(0.041)	(0.160)	(0.673)	(0.037)	(0.163)	(0.162)	(0.033)	(0.162)	(0.161)
RATING3	1.400	0.576	0.673	1.787	0.663	0.739	1.673	0.611	0.707
	(0.040)	(0.154)	(0.153)	(0.036)	(0.157)	(0.156)	(0.032)	(0.156)	(0.155)
RATING4	1.846	0.795	0.881	2.299	0.956	1.023	2.240	0.939	1.026
	(0.040)	(0.154)	(0.153)	(0.036)	(0.156)	(0.156)	(0.032)	(0.155)	(0.156)
Group level									
PRICE2			0.301			0.268			0.217
			(0.068)			(0.068)			(0.068)
PRICE3			0.727			0.603			0.454
			(0.160)			(0.155)			(0.152)
REVIEWS			-0.408			-0.394			-0.344
			(0.059)			(0.060)			(0.060)
LNCOMP			0.479			0.458			0.375
			(0.089)			(0.089)			(0.088)
INEQ			-0.171			-0.118			-0.184
			(0.170)			(0.168)			(0.166)
LNINC			0.115			0.010			0.055
			(0.067)			(0.066)			(0.066)
LNPOP			-0.064			-0.053			-0.067
			(0.050)			(0.050)			(0.049)
CONST		-0.631	0.003		-0.765	-0.012		-0.771	0.713
		(0.153)	(1.004)		(0.156)	(0.995)		(0.155)	(0.987)
No. of markets	236	236	236	239	239	239	238	238	238
LogL	-188,783	-15,740	-15,696	-310,566	-17,289	-17,253	-416,735	-18,144	-18,113
Pearson	443,831	3,559	3,522	742,451	3,450	3,426	983,624	3,352	3,338
AIC	377,573	31,487	31,414	621,137	34,587	34,528	833,475	36,296	36,315

Source: Yelp. Dependent variable: Cumulative checkins of individual restaurants. Standard errors in parentheses. Markets are defined on a ZIPxPRICE level. Dirmul I refers to $\delta_g = \delta$, Dirmul II to $\delta_g = f(x_g) = \exp(-\gamma' w_g)$. *RATING* dummy variables reflect a 4-step scale. *PRICE* dummy variables reflect three price categories. *REVIEWS* is the logarithm of the sum of reviews in a market. *LNCOMP* is the logarithm of the number of competitors in a market. *INEQ* is the ratio of average income and median income in a ZIP code area, proxying income inequality. *LNINC* is the logarithm of average income in the ZIP code area. *LNPOP* is the logarithm of the ZIP population.

Table 3.3: Expectations and correlations across time

	2010-2012		2013		2014	
Variable	*MNL*	*Dir I*	*MNL*	*Dir I*	*MNL*	*Dir I*
Cumulative checkins						
δ		1.879		2.149		2.162
ρ		0.233		0.239		0.245
$\delta^{-1}\alpha 1$		0.532		0.465		0.462
$\delta^{-1}\alpha 2$		0.636		0.614		0.582
$\delta^{-1}\alpha 3$		0.947		0.903		0.852
$\delta^{-1}\alpha 4$		1.179		1.210		1.183
Single period checkins						
δ				2.868		3.090
ρ				0.336		0.351
$\delta^{-1}\alpha 1$				0.349		0.324
$\delta^{-1}\alpha 2$				0.355		0.331
$\delta^{-1}\alpha 3$				0.512		0.477
$\delta^{-1}\alpha 4$				0.758		0.719

Source: Yelp. The upper half of the table shows results for cumulative checkins, the lower half shows results for single-period checkins in years 2013 and 2014. Dirmul I refers to $\delta_g = \delta = \exp(-constant)$. Calculations assume that exactly one restaurant of each rating category exist within a group and are based on the results in Table 3.2 and 3.5. Bottom half uses L1-lags for 2013 and L2-lags for 2014. δ measures the strength of interactions, with $\delta \to 0$ indicating no social interactions, while $\delta \to \infty$ presents a the limit case where visits depend exclusively on other individuals' previous choices. $\rho = 1/(1 + \sum_{j=1}^{R} \delta^{-1}\alpha_j)$ is the overdispersion parameter converging to 0 as $\delta \to 0$.

Table 3.4: Cumulative checkins over time, aggregated by rating

	2010-2012			2013			2014		
Variable	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir I*	*Dir II*
RATING2	0.946	0.354	0.505	1.502	0.444	0.566	1.306	0.433	0.588
	(0.041)	(0.189)	(0.188)	(0.037)	(0.186)	(0.186)	(0.033)	(0.184)	(0.185)
RATING3	3.522	1.640	1.908	4.216	1.787	2.022	4.058	1.742	1.992
	(0.040)	(0.184)	(0.188)	(0.036)	(0.181)	(0.188)	(0.032)	(0.179)	(0.185)
RATING4	3.723	1.895	2.127	4.403	1.990	2.199	4.295	1.948	2.127
	(0.040)	(0.183)	(0.188)	(0.036)	(0.181)	(0.188)	(0.032)	(0.179)	(0.184)
Group level									
PRICE2			0.003			-0.104			0.057
			(0.177)			(0.176)			(0.168)
PRICE3			0.193			0.137			0.087
			(0.342)			(0.349)			(0.320)
REVIEWS			-0.409			-0.323			-0.447
			(0.128)			(0.132)			(0.129)
LNCOMP			0.911			0.760			0.811
			(0.214)			(0.216)			(0.209)
INEQ			-0.446			-0.292			0.016
			(0.514)			(0.529)			(0.511)
LNINC			0.002			0.043			0.158
			(0.188)			(0.196)			(0.183)
LNPOP			-0.057			-0.088			0.094
			(0.127)			(0.136)			(0.126)
CONST		- 0.948	0.160		-1.091	-0.260		-1.132	-3.191
		(0.176)	(2.698)		(0.172)	(2.958)		(0.169)	(2.796)
No. of observations	592	592	592	598	598	598	593	593	593
No. of markets	220	220	220	223	223	223	221	221	221
LogL	-37,723	-2,172	-2,157	-63,836	-2,377	-2,365	-89,520	-2,480	-2,468
Pearson	74,659	354	372	132,279	370	382	183,405	342	355
AIC	75,451	4,352	4,335	127,677	4,762	4,751	179,046	4,969	4,957

Source: Yelp. Dependent variable: Cumulative checkins, aggregated by rating within a market. Standard errors in parentheses. Markets are defined on a ZIPxPRICE level. Dirmul I refers to $\delta_g = \delta$, Dirmul II to $\delta_g = f(x_g) = \exp(-\gamma' w_g)$. $RATING$ dummy variables reflect a 4-step scale. $PRICE$ dummy variables reflect three price categories. $REVIEWS$ is the logarithm of the sum of reviews in a market. $LNCOMP$ is the logarithm of the number of competitors in a market. $INEQ$ is the ratio of average income and median income in a ZIP code area, proxying income inequality. $LNINC$ is the logarithm of average income in the ZIP code area. $LNPOP$ is the logarithm of the ZIP population.

Table 3.5: Single period checkins over time

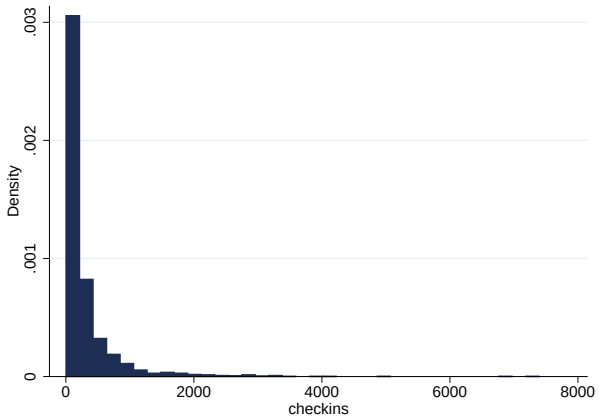
	2013			2014					
Variable	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir II*	*Dir II*
LRATING2	0.524	0.017	0.062	0.912	0.217	0.241			
	(0.051)	(0.182)	(0.182)	(0.072)	(0.208)	(0.208)			
LRATING3	1.470	0.384	0.466	1.939	0.617	0.687			
	(0.050)	(0.174)	(0.174)	(0.070)	(0.200)	(0.200)			
LRATING4	2.052	0.776	0.853	2.666	1.048	1.110			
	(0.050)	(0.174)	(0.174)	(0.070)	(0.200)	(0.201)			
L2RATING2							0.603	0.022	0.076
							(0.040)	(0.183)	(0.182)
L2RATING3							1.558	0.387	0.481
							(0.039)	(0.175)	(0.175)
L2RATING4							2.174	0.798	0.886
							(0.039)	(0.175)	(0.175)
Group level									
PRICE2			0.025			0.042			0.047
			(0.070)			(0.026)			(0.068)
PRICE3			0.112			-0.015			0.059
			(0.163)			(0.052)			(0.158)
LREVIEWS			-0.178			-0.148			
			(0.061)			(0.064)			
LCOMP			0.140			0.025			
			(0.092)			(0.095)			
L2REVIEWS									-0.171
									(0.059)
L2COMP									0.108
									(0.089)
INEQ			-0.312			-0.317			-0.257
			(0.181)			(0.190)			(0.179)
LNINC			0.027			-0.080			-0.041
			(0.071)			(0.072)			(0.069)
LNPOP			-0.067			-0.029			-0.061
			(0.053)			(0.055)			(0.052)
CONST		-1.054	0.440		-1.521	0.993		-1.128	1.027
		(0.173)	(1.041)		(0.198)	(1.078)		(0.173)	(1.014)
No. of markets	237	237	237	238	238	238	236	236	236
LogL	-147,888	-14,139	-14,111	-134,399	-13,390	-13,359	-268,138	-15,753	-15,626
Pearson	327,815	2,427	2,448	294,048	2,151	2,173	605,410	2,311	2,326
AIC	295,781	28,270	28,245	268,805	26,788	26,740	536,283	31,514	31,341

Source: Yelp. Dependent variable: Single-period checkins of individual restaurants. Standard errors in parentheses. Markets are defined on a ZIPxPRICE level. Dirmul I refers to $\delta_g = \delta$, Dirmul II to $\delta_g = f(x_g) = \exp(-\gamma' w_g)$. $RATING$ dummy variables reflect a 4-step scale. $PRICE$ dummy variables reflect three price categories. $REVIEWS$ is the logarithm of the sum of reviews in a market. $LNCOMP$ is the logarithm of the number of competitors in a market. $INEQ$ is the ratio of average income and median income in a ZIP code area, proxying income inequality. $LNINC$ is the logarithm of average income in the ZIP code area. $LNPOP$ is the logarithm of the ZIP population.

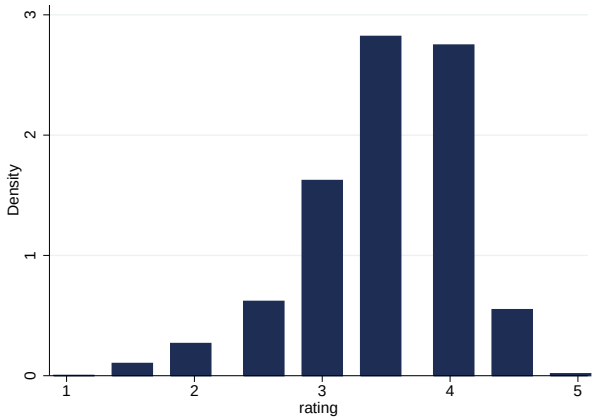
Table 3.6: Cumulative checkins over time, higher grouping levels

	ZIP, 2014			City X Price, 2014			City, 2014		
Variable	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir I*	*Dir II*	*MNL*	*Dir II*	*Dir II*
RATING2	0.689	0.219	0.280	0.677	0.239	0.316	0.675	0.209	0.281
	(0.032)	(0.160)	(0.159)	(0.032)	(0.157)	(0.156)	(0.032)	(0.158)	(0.158)
RATING3	1.646	0.543	0.622	1.752	0.604	0.701	1.685	0.526	0.612
	(0.031)	(0.154)	(0.154)	(0.031)	(0.150)	(0.150)	(0.031)	(0.153)	(0.153)
RATING4	2.139	0.817	0.889	2.313	0.921	1.011	2.174	0.780	0.856
	(0.031)	(0.159)	(0.158)	(0.031)	(0.151)	(0.150)	(0.031)	(0.158)	(0.158)
PRICE2	1.010	0.123	1.024				4.554	3.753	1.001
	(0.082)	(0.798)	(0.949)				(0.127)	(1.294)	(1.660)
PRICE3	-0.519	1.134	2.394				5.877	5.355	3.387
	(0.190)	(2.032)	(2.115)				(0.267)	(3.116)	(3.278)
PRICERAT2	0.152	0.122	0.133				0.148	0.133	0.129
	(0.004)	(0.047)	(0.046)				(0.004)	(0.047)	(0.046)
PRICERAT3	-0.086	0.002	0.027				0.075	0.082	-0.316
	(0.013)	(0.124)	(0.122)				(0.013)	(0.126)	(0.295)
PRICEINC2	-0.074	-0.017	-0.101				-0.384	-0.343	-0.096
	(0.007)	(0.070)	(0.084)				(0.011)	(0.116)	(0.149)
PRICEINC3	0.084	-0.088	-0.206				-0.534	-0.488	-0.316
	(0.016)	(0.177)	(0.186)				(0.023)	(0.279)	(0.295)
Group level									
PRICE2						0.240			
						(0.116)			
PRICE3						0.414			
						(0.208)			
REVIEWS			-0.349			-0.418			-0.548
			(0.064)			(0.126)			(0.224)
LNCOMP			0.385			0.353			0.565
			(0.097)			(0.154)			(0.294)
INEQ			-0.182			0.049			0.203
			(0.164)			(0.184)			(0.249)
LNINC			0.168			-0.164			0.142
			(0.078)			(0.149)			(0.292)
LNPOP			-0.066			0.045			-0.008
			(0.049)			(0.058)			(0.081)
CONST		-0.909	-0.543		-0.859	2.092		-0.992	-0.464
		(0.154)	(1.087)		(0.149)	(1.793)		(0.153)	(2.989)
No. of markets	118	118	118	71	71	71	37	37	37
LogL	-439,340	-19,310	-19,277	-514,537	-20,046	-20,016	-520,212	-20,488	-20,462
Pearson	1,085,552	3,630	3,610	1,413,224	4,288	4,271	1,433,497	4,371	4,350
AIC	878,697	38,641	38,584	1,029,080	40,099	40,054	1,040,443	40,997	40,954

Source: Yelp. Dependent variable: Cumulative checkins of individual restaurants. Standard errors in parentheses. Markets are defined on a ZIP level. Dirmul I refers to $\delta_g = \delta$, Dirmul II to $\delta_g = f(x_g) = \exp(-\gamma' w_g)$. $RATING$ dummy variables reflect a 4-step scale. $PRICE$ dummy variables reflect three price categories. $REVIEWS$ is the logarithm of the sum of reviews in a market. $LNCOMP$ is the logarithm of the number of competitors in a market. $INEQ$ is the ratio of average income and median income in a ZIP code area, proxying income inequality. $LNINC$ is the logarithm of average income in the ZIP code area. $LNPOP$ is the logarithm of the ZIP population.

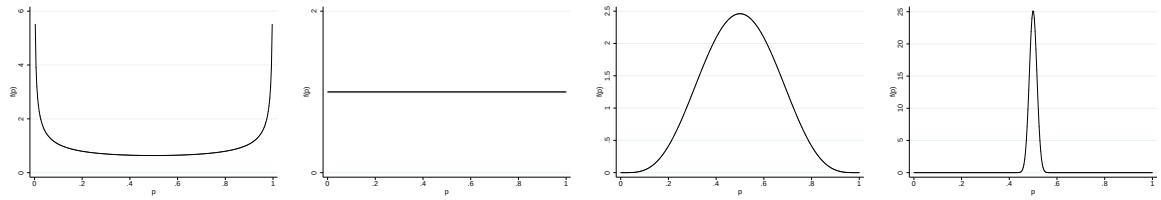


(a) Checkins distribution as of 2015



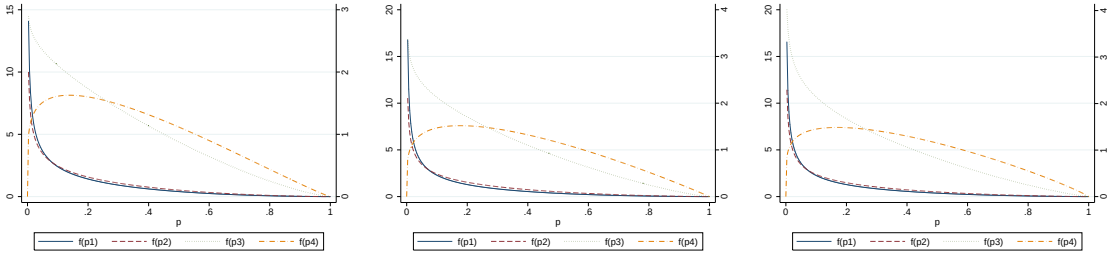
(b) Ratings distribution as of 2015

Figure 3.1: Restaurant visits and ratings as of 2015



(a) $\delta = 2.000$ (b) $\delta = 1.000$ (c) $\delta = 0.200$ (d) $\delta = 0.002$

Figure 3.2: Dirichlet distributions for a set of different δ



(a) 2010-2012 (b) 2013 (c) 2014

Figure 3.3: Dirichlet parameters for different rating categories, over time

Chapter 4

Predicting US bank failures with internet search volume data

This chapter is a revised version of Working Paper No. 214 published in the *Working Paper Series* of the Department of Economics, University of Zurich.

Acknowledgements: I thank Michel Habib, Steven Ongena, Rainer Winkelmann, Raphael Studer, two anonymous referees at the Journal of Banking and Finance as well as the participants of the Zurich Workshop on Economics 2013 for helpful comments and suggestions.

4.1 Introduction

When US-Senator Chuck Schumer publicly questioned the financial health of the bank IndyMac in the summer of 2008, the bank's customers were quick to react. Within just three days, IndyMac lost USD 100 million in deposits (Los Angeles Times, 2008); after thirteen days and withdrawals amounting to USD 1.3 billion, the bank failed (Grind, 2012; Seabrook, 2008). These developments were closely tracked by the Google search volume index; on the day of Schumer's announcement, the index value for the search term "IndyMac" almost doubled, rising further up to 22-fold in the following days. During the same period Washington Mutual, a struggling competitor, lost a total of USD 9.4 billion due to a bank run, even surpassing IndyMac's deposit loss. Contrary to the much publicized IndyMac incident, the Washington Mutual run was largely unnoticed by the media or the analysts (Grind, 2012). The run did not escape the Google search volumes and Washington Mutuals share price, however: As in the IndyMac case, the search volume index for "Washington Mutual" more than doubled during the days of the run. The peak of the bank run on Tuesday, July 14, 2008 coincided with the high search volume on that same day. Two months later, Washington Mutual went through another bank run, with its peak on Thursday, September 18. Again, Google index values track a total outflow of approximately USD 16.7 billion between September 15 and September 24 (Office of Thrift Supervision, 2008).[1] Search volumes were surpassed only on September 25, 2008, the day when the Federal Deposit Insurance Company (FDIC) walked into Washington Mutual's offices and shut the bank down (see Figure 4.1).[2]

These two examples provide anecdotal evidence that Google search volumes can be a valuable proxy to reflect public attention, which is generally hard to capture. It has been shown in a variety of settings that Google can be instrumented to reflect such phenomena,

[1] Although there was extensive media coverage on the bank's health before its closure, the public was informed about the bank run only after the fact (Grind, 2012).

[2] The low index values occurring in regular intervals are typically weekend days - days when individuals spend less time on their computers and bank transactions cannot be executed.

from influenza epidemics to unemployment forecasting (Ginsberg et al., 2008; Breyer et al., 2011; McCarthy, 2010; Bollen et al., 2011). Whether it is useful for tracking developments in the banking industry is the subject of this paper.

Contrary to, say, a newspaper article, internet search volumes reflect a much more crowd-sourced and democratic approach. Users are actively looking for something rather than consuming information passively. While a newspaper article about a specific bank must be considered worthwile reporting in the first place and, in a second step, restricted to reflecting the facts, Google search queries can capture a much wider set of information: facts as well as speculations, banking experts as well as individual savers. Google queries also give an idea of how many individuals care about a specific topic. Each search query is an uptick in the volume, translated into a rising index value. In that sense, it is similar to trading volumes in financial markets (see also Mathiesen et al. 2013) - except that for the majority of the banks considered in this paper, such volumes do not exist because many of them are not listed on a stock exchange. In the absence of share prices, Google data may therefore prove to be a valuable source for understanding and predicting bank failures. Providing real-time data on the popularity of a bank's name on the internet on a weekly basis, Google can help modelling short term dynamics by incorporating information that is not fully captured in balance sheet positions or macro-level variables. Such information might be important: The deposit withdrawal at Washington Mutual in July 2008 is what Iyer et al. (2013) call a non-fundamental shock: a run that cannot be justified by the balance sheet fundamentals of the bank itself or that couldn't have already been justified at an earlier point in time. At the time such a shock would have become visible in Washington Mutual's balance sheet, the bank was already under the reign of JP Morgan. Google tracked that non-fundamental shock in a timely manner. Google data is also an interesting addition to fundamentals in light of the Iyer et al. (2013) finding that large depositors tend to orient themselves to and act on (possibly non-public) regulatory actions rather than fundamentals; Google might partly capture these movements.

Three main questions are tackled in this paper. Since Google data is not available for all banks (discussed in section 4.3.1), a first question seeks to answer whether the availability of Google data itself correlates with the survival of an individual bank. Second, given that Google data is available, the question of how well the Google query shares track bank failures is examined. Thirdly, I discuss the question of how indicative past changes in search volumes are when trying to predict future failures.

To answer these questions, this paper looks at 433 bank failures and 400 surviving banks in the United States in the period from January 2007 to March 2012, working with a dataset including both Google data as well as balance sheet and revenue data on the level of an individual bank. Using an exponential duration model with a piecewise-constant hazard rate and time-varying covariates, I analyze how well Google search volumes in the United States track and predict these bank failures. Results show that while the availabilty of Google data itself has no significant effect on a bank's survival, higher Google search volumes go hand in hand with higher hazard rates. As one moves further away from the failure date, Google's predictive power dissipates.

The paper is structured as follows: in the following section, previous findings are discussed. In the third section, the data is presented. In the fourth section, I model the failure rates of individual banks, using weekly Google time series and balance sheet positions and revenue data from the FDIC as explanatory variables. Section 5 concludes.

4.2 Previous literature

Working with Google data to model short-term developments has been successful before. From influenza epidemics (Ginsberg et al., 2008) to tracking kidney stone incidences (Breyer et al., 2011) and monitoring suicide risks (McCarthy, 2010), on to more economic applications in the field of unemployment (Askitas and Zimmermann, 2009; D'Amuri and Marcucci, 2010; Choi, 2009; Tefft, 2011), inflation (Guzman, 2011), consumer behavior (Choi and Var-

ian, 2012; Goel et al., 2010), consumer sentiment (Radinsky et al., 2008; Della Penna and Huang, 2009; Preis et al., 2010) and housing prices (McLaren and Shanbhogue, 2011; Wu and Brynjolfsson, 2013). Financial markets have received some attention, too: Preis et al. (2013) quantified trading behavior using Google, Bollen et al. (2011) predict stock market movements using Twitter, Mathiesen et al. (2013) likened the statistical properties of Twitter data to the properties of trading volumes of stocks and Moat et al. (2013) studied the correlation of Wikipedia page views and stock market movements. To my knowledge, no paper to date has used Google search volumes to predict bank failures.

While there has been a variety of empirical work studying both wider banking panics as well as individual bank failures, this literature has focussed on balance sheet positions and revenue data, looking at issues of panics, contagion and information networks (for an overview, see Gorton and Winton, 2003). Calomiris and Mason (2003) analyze bank failures in the 1920's and 1930's using a duration model and data on individual banks as well as regional economic factors, disputing the Friedman-Schwartz argument that many bank failures resulted from unwarranted panic and finding evidence that most of the failures are justified by weak fundamentals. Saunders and Wilson (1996) look at the role of bank contagion and information in the same period, using data on deposit flows. Wheelock and Wilson (2000) make use of duration models to determine the effect of managerial inefficiency on the probability of failure and acquisition. Whalen (1991) assesses the usefulness of using proportional hazard models as early warning tool, concluding that "reasonably accurate early warning models can be built and maintained at relatively low cost."

Short term dynamics and irrational elements leading bank failures have proven difficult to account for. Regarding bank runs on individual banks and micro-level withdrawal patterns, there exists only a small literature. A recent one is Iyer and Puria (2012), which looks at the dynamics of withdrawal patterns, deposit insurance and social networks in an Indian bank. A follow-up study (Iyer et al., 2013) looks at how depositors monitor banks, finding that regulatory agencies play an important role in the monitoring process. Other

examples in the area of individual failures and information networks include Kelly and O Grada (2000) or O Grada and White (2003).

From a macro perspective, Donaldson (1992) finds that there are periods when banking panics are more likely to occur, but that exact starting dates of such panics are unpredictable. Gorton (1988) offers empirical evidence compatible with the idea that when depositors receive information forecasting a recession, they draw on their bank accounts, knowing that they will be dissaving and anticipating the higher bank failure rate during recessions. I try to take such factors into account by including macro-level variables.

4.3 Data sources, data properties and descriptive statistics

As of December 2006, there were 8,681 active banks insured by the FDIC, compared to 7,357 at the end of 2011. Within these five years, 433 FDIC-insured banks failed.[3] In a first sample, I include the 433 failed banks in the period from January 1, 2007 to March 31, 2012. In addition, I randomly select a subset of 400 banks from the set of 7,357 active banks at the end of 2011 to include in the sample as control observations.[4] Focussing on a random sample of surviving banks instead of using the full sample is a result of the data collection procedure: As each query on Google needs to be executed manually, collecting data on the whole set of surviving banks is infeasible.

To restore adequate proportions between failures and survivors, I weight observations accordingly when estimating the models (discussed in Section 4.4). To have an equal entering date for all banks at risk of failure and to avoid complications when weighting observations (also discussed in Section 4.4), 18 banks founded after January 1st, 2007 were

[3]Note that aside from failures, there also were mergers as well as newly founded banks.

[4]None of these randomly selected banks were merged into other banks or failed up to the first quarter of 2012. Sampling was done at the end of 2011 rather than at the end of the first quarter 2012 since data on the first quarter of 2012 was only added at a later stage.

dropped, of which 5 were failed banks. For the resulting 815 banks, weekly Google search queries data and quarterly FDIC data was downloaded. The sample period covers 273 weeks or 21 quarters.

4.3.1 Data sources: Google Insights for Search

Weekly search query time series containing the bank's name have been executed and downloaded on "Google Insights for Search" (Google, 2012), Google's tool to analyze search volumes.[5] These time series reflect the query share of the bank's name in the overall search traffic categorized as "Finance" on a weekly basis. The structure and properties of Google data deserves some extra attention, as it has some non-standard restriction features.

The first restriction concerns the time horizon: Google time series go no further back than January 1, 2004 (Choi and Varian, 2012). There is no data available before that date.

Second, Google only publishes relative numbers, not absolute search volume numbers. The numbers are relative in two dimensions. First, the query share QS_{ijut} is the ratio of the number of queries n_{ijut} for a given search term i and the total number of queries N_{jut} in the selected category j in geographic area u at time t :

$$QS_{ijut} = \frac{n_{ijut}}{N_{jut}}, \quad 0 \leq QS_{ijut} \leq 1$$

The second dimension concerns the time series of the query share itself. All query shares are reported relative to the maximum query share M_{ijuS} in the selected period S multiplied by 100, which gives the Google index value GI_{ijut}:

$$GI_{ijut} = \frac{QS_{ijut}}{M_{ijuS}} \times 100,$$

$$\text{with} \quad M_{ijuS} = \max_{t \in S} QS_{ijut}, \quad 0 \leq GI_{ijut} \leq 100$$

GI_{ijut} is the number published by Google; all other numbers are not published. Under the assumption that internet usage is growing, a rising index value can always be interpreted

[5] "Google Insights for Search" has been renamed to "Google Trends" in the meantime.

as a rise in popularity for the search term. This is not true for falling values, as it is enough for the search term to be growing at a less than average rate in order for the index value to fall. Growth rates in query shares from t to $t+1$ are preserved in the published relative numbers, whereas percentage point differences are not. The levels of the index values are not comparable across banks. For these reasons, only Google growth rates are used in this paper.

Third, queries are "broad matched", meaning that queries such as "IndyMac bank run" are counted in the calculation of the query index for "IndyMac", but not vice-versa. Entering less and more general search terms increases the probability that unrelated searches are captured as well. For example, a query with the search term "forecast" may capture results related to forecasts of economic indicators, election results or weather, whereas a query for "weather forecast tomorrow Zurich" is much more specific and unlikely to include unrelated queries.

Fourth and linked to the third restriction, Google series for more restrictive queries are more likely not to be published at all. As mentioned above, Google publishes the index values only if the absolute number of search queries exceeds an unknown threshold (Choi and Varian, 2012). This has two consequences: First, it restricts the sample from 815 to 210 banks for which any Google data is available. Second, within the remaining 210 banks, the absolute search volume might temporarily fall under the threshold and a value of zero is published. Since the true index value is greater than or equal to zero, using these time series can bias estimation results. Focusing only on the complete cases with uncensored Google series, on the other hand, reduces the population to 25 failing banks and 23 surviving banks.[6]

Google data is retrieved for the period of January 4, 2004 to March 31, 2012. Time series are on a weekly basis. Queries are restricted to the United States and to the "Finance"

[6]When calculating percentage changes for a Google series that was censored in the preceding period, the value was set to missing.

category to avoid counting unrelated queries in the index.[7] Queries outside the United States are unlikely to be related to the individual banks, while narrowing the geographic space to state levels would have resulted in more censored time series. A similar logic applies to the categoric restriction to "Finance": With a broader definition, unrelated queries might be captured in the index, while a narrower definition might exclude relevant queries or leads to censoring.

For each bank, there is a separate Google query containing the bank's name as a search term. For practical reasons, legal appendices such as *FSB*, *NA*, *National Association* or *Company* as well as "The" and "&" in bank names have been dropped, as it is unlikely that individuals search for their bank with legal appendices or include symbols such as "&".[8] Likewise, missing spaces (such as in *WashingtonFirst*) have been inserted. As for the case of popular bank names, there are three institutions named "First State Bank", two "The First State Bank", two "Premier Bank", two "Summit Bank", two "The Park Avenue Bank", two "Legacy Bank", two "First National Bank", two "Citizens National Bank" and two "Integrity Bank" in the sample.[9] In these cases, identical Google query time series have to be used, as one cannot differentiate and assign unique series to each institution.

4.3.2 Additional data sources

The second major data souce for this paper is the FDIC database (Federal Deposit Insurance Corp., 2011). The FDIC provides a large set of balance sheet positions, revenue figures and other characteristics of individual banks, which a number of researchers have used for similar estimations. For the purpose of this paper, 11 variables were selected and downloaded on the bank level on a quarterly basis, the shortest time interval available.

[7] Google classifies queries into about 30 categories at the top level and about 250 categories at the second level using a natural language classification engine (Choi and Varian, 2012).

[8] The exact names used for the queries are stored in the *Search_name* variable - a missing value means that the name has been used without any modification.

[9] Note that the "The" in bank names was dropped when Google data was downloaded, i.e. effectively there are five banks named "First State Bank".

The variables can be broadly classified in the categories capital adequacy, asset quality, earnings, liquidity and other factors. The selection of the variables was guided by the selections in previous research papers estimating similar models (e.g. Cole and Gunther 1995; Calomiris and Mason 2003; Wheelock and Wilson 2000). In addition, an indicator variable for the FDIC insurance limit raise from USD 100,000 to USD 250,000 on October 3, 2008 was defined. The FDIC dataset comprises 836 institutions.

Bloomberg serves as an additional data source from which weekly LIBOR and overnight indexed swap (OIS) time series were downloaded. 2010 US Census data (United States Census Bureau, 2010) was used to define urban area dummy variables on the US county level.

4.3.3 Variables and summary statistics

Information on bank failures is taken from the FDIC, which lists failures in its failed banks list. The failure date is defined as the closing date that the FDIC lists on that same list. The FDIC has some discretion when it comes to the exact date of the closing, and therefore to exploit the weekend days to wind down a bank (i.e. when banks are closed), most of these closing dates are on a Friday. For my purposes, this means I can aggregate these closing dates to a weekly measure with little loss of information. The failure time is then defined as the week into which the closing date falls. Table 4.2 lists the number of failures in a given year. As one can see from the table, most of the failures occur in the years after 2006. With respect to survival analysis, there is little information in the years 2004 to 2006 since there are almost no failures. In addition, these failures are unlikely to be connected to the financial crises. I therefore dropped the years 2004 to 2006.[10]

Summary statistics are presented in Table 4.3. Aside from the Google variable, several balance sheet variables are listed, which can be roughly categorized into a capitalisation

[10]I did run the analysis including these years as a robustness check, without any meaningful changes in the results.

variable (capital), asset quality variables (troubled assets, commercial real estate, residential real estate), earnings (net income), liquidity (large CDs, insdep, securities) and miscellaneous factors (insider loans, holding company, entering age, urban). A description of the variables can be found in Table C.1 in the appendix. The reported values in the table are averages over the period starting in January 2007 to March 2012 or the respective failure date where in a first step, the average over all periods is taken for each bank, and then the average is taken over all banks in the group (i.e. failures/survivors). The upper third includes all banks, the middle third only banks where Google data is available, and the bottom third only banks with uncensored Google series.

Even if averaged over time, survivors and failures differ in some of the variables, as can be seen by the stars indicating a difference in the Wilcoxon rank-sum test at the one percent significance level. Differences attenuate somewhat as one restricts the sample to banks having uncensored Google series, which is the sample main results will be based on. The differences in capital ratio or large cash deposits, for example, are not significant anymore. If you exclude the two largest failures, Washington Mutual and IndyMac, from the sample, the difference in gross assets is not significant anymore either. Surviving banks differ from their failing counterparts with respect to troubled assets, net income, securities and age.

In terms of changes in Google query shares during the last weeks prior to failure, other bank failures resemble the pattern of Washington Mutual seen in the introduction. Figure 4.2 shows the weekly mean of the growth rate of Google query shares for the names of the 25 failed banks with uncensored Google series, compared to the corresponding means of the 23 surviving banks. To calculate the value for the control group, control group values were averaged in the corresponding week to failure for the failing bank. In a second step, these values were averaged over all failing banks. Values for failing banks remain on a low level up to five weeks before failure. From then on, there is a slight upward trend in rates, up to about one week before failure, when they spike and remain high in the weeks of and

after the failure. Shortly after the failure, rates drop sharply. Meanwhile, changes in query shares for surviving banks stay constant.

Figure 4.3 shows the quarterly means of different key balance sheet positions of failed banks in the last quarters before failure, again contrasted by the same statistics for surviving banks (the values were calculated analogously to the Google values above). Note that the horizontal axis is measured in quarters as opposed to weeks in the graph before. One can see clear trends in capital ratios and troubled assets ratios that start out at least one year before failure. Ratios for large deposits and securities stay relatively constant over time, but show clear differences across the failing and the surviving group. Comparing these graphs suggests that fundamentals of failing banks deteriorate early, while surviving banks' advantageous securities and large deposit positions protect them when having to react to liquidity drains. Google search queries, on the other hand, react when failure is imminent, correlating with the timing of failure rather than with the probability of failure itself.

4.4 Model and results

4.4.1 Model

I use a piecewise-constant exponential model to model bank failures, estimating the hazard rate semi-parametrically. Using a piecewise-constant hazard as opposed to a parametric model such as the exponential or Weibull has the advantage of modelling the baseline hazard semi-parametrically. This is important as the baseline hazard, i.e. the hazard common to all banks, is likely to change over time, especially during the financial crisis. To account for the changes in the hazard rate over time and work with time-varying covariates, the dataset is split into 273 weekly episodes.

The baseline hazard is modelled using time dummies as well as macro-variables (the LOIS spread). With respect to time dummies, three specifications will be used. The first

involves splitting the 2007 to 2012 into just two subperiods: one before and one after the raise of the FDIC insurance limit from USD 100,000 to USD 250,000 in October 2008. This intervention is mainly an intervention to prevent potential bank runs from depositors; whether the hazard changes can be directly tested on the corresponding dummy variable for the intervention. A second specification uses yearly time dummies, changing the baseline hazard every year. A third specification uses quarterly dummies. As an alternative to the piecewise-constant hazard model and as a robustness check, I also estimate a Cox proportional hazard model. Note that in this case, the baseline hazard function is completely unspecified.

In the piecewise-constant hazard model, the hazard rate is a step function specified as

$$\begin{aligned}\theta(t, \mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t) &= \theta_0(t)\lambda(\mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t) \\ &= \bar{\theta}_t \exp(\beta'\mathbf{x}_{it} + \delta'\mathbf{z}_i + \gamma'\mathbf{w}_t) \\ &= \exp[\log(\bar{\theta}_t) + \beta'\mathbf{x}_{it} + \delta'\mathbf{z}_i + \gamma'\mathbf{w}_t] \\ &= \exp(\tilde{\lambda}_t)\end{aligned}$$

where $\bar{\theta}_t$ is the interval-specific baseline hazard common to all banks and $\lambda(\mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t)$ is the bank-specific hazard component in period t. $\mathbf{x}_{it}$ is a vector including individual time-varying covariates, $\mathbf{z}_i$ contains individual time-constant covariates and $\mathbf{w}_t$ contains common, time-varying elements at time t. The interval-specific baseline hazard is equivalent to including a period-specific dummy variables in the overall hazard.

In the case of two subperiods with $\mathbf{x}_{it} = \mathbf{x}_{i1}$ and $\mathbf{w}_t = \mathbf{w}_1$ if $t < s$ and $\mathbf{x}_{it} = \mathbf{x}_{i2}$ and $\mathbf{w}_t = \mathbf{w}_2$ if $t \geq s$, the corresponding survivor function is given by (see Jenkins, 2005)

$$\begin{aligned}S(t, \mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t) &= [S_0(s)]^{\tilde{\lambda}_1} \frac{[S_0(t)]^{\tilde{\lambda}_2}}{[S_0(s)]^{\tilde{\lambda}_2}} \\ &= \exp(-s\tilde{\lambda}_1) \exp\left[-(t-s)\tilde{\lambda}_2\right]\end{aligned}$$

Note that Google data varies on a weekly basis, while balance sheet data varies only

quarterly.

As mentioned in the data section, the sample consists of all banks that failed in the period from January 2007 to March 2012, plus a random sample of surviving banks. While in the sample of 815 banks, more than half of them fail within the roughly eight years covered, these failures represent only about five percent of the whole bank population of 8,681 institutions in December 2006. This choice-based sampling therefore needs to be accounted for by weighting observations accordingly (Lancaster, 1992). Individual likelihood contributions are weighted by P/Q, where P represents the population fraction of failing institutions, and Q represents the sample fraction of failing institutions (correspondingly, $(1-P)$ and $(1-Q)$ are the weights for surviving institutions). Accordingly, failing banks get a lower weight than surviving banks. I reported both the absolute number of failures as well as the weighted failures in the result tables.

4.4.2 Results

In this section, I seek to answer the three questions raised in the introduction. The empirical hazard rate including all 815 banks is displayed in Figure 4.4. One can see that the hazard rate varies with time, peaking after week 170, or at the beginning of 2010. The result tables are split into five columns; the model in the first column is using Google as the only explanatory variable. The second column shows results for a piecewise-constant hazard rate model with two subperiods (pre- and post FDIC insurance limit raise), followed by the piecewise-constant hazard models with yearly and quarterly dummies. The last column shows the results of the Cox proportional hazards model. The bottom of the table specifies whether the Google data availability variable (Dummy) or the Google growth variable for the percentage change from the last period (Growth) was used. The table also lists the total number of banks, the absolute number of failures, and the weighted number of failures (which is around five percent of the total number of banks, as outlined in the previous section). Note that the majority of the explanatory variables are roughly

bounded between 0 and 100, as they are percentages of gross assets. The tables report coefficients (as opposed to hazard ratios). A change in X_k changes the overall hazard by $\frac{\partial\theta(t,\mathbf{X_t},\mathbf{Z},\mathbf{W_t})}{\partial X_{kt}} = \theta(t,\mathbf{X_t},\mathbf{Z},\mathbf{W_t})\beta_k$ or increases the hazard by $100(\exp(\beta_k)-1)$ percent (approximatly $100\beta_k$ percent). A negative coefficient decreases the hazard accordingly.

Table 4.4 shows results using the Google dummy variable. The coefficient for the Google variable is positive, but remains statistically insignificant in all five models - whether Google data is available is not a significant predictor whether a bank fails or not. Capital has the anticipated, large negative effect on the hazard rate. Troubled assets positions increase the hazard rate as one would expect, while securities - which can serve as collateral when lending money - decrease the hazard, as do large cash deposits. Interestingly, the coefficient for the variable *insdep*, the interaction between the large deposits ratio and the FDIC insurance limit raise dummy, is positive, implying a relatively higher hazard for banks with large deposits after the FDIC intervention. Note that the coefficient on the FDIC intervention (the subperiod dummy) counteracts the effect with a negative coefficient of about the same magnitude (not shown in the table). Finally, coefficients on commercial real estate and *urban* are positive, while the coefficient for holding companies is negative. The remaining effects are not statistically significant.

Table 4.5 presents the main results using only banks with uncensored Google series. Generally, effects increase the more flexible the baseline hazard is specified, with the exception of the macro-variable LOIS, whose effect is increasingly captured by the more flexible baseline hazard time dummies as one moves from the left to the right of the table. Coefficients are in line with expectations. In all specifications, the coefficient on the Google variable is significant, raising the hazard rate between approximately 2.4 and 4.8 percent, which is roughly comparable to the coefficient on troubled assets. Capital and securities have the largest effects, both reducing the hazard rate. The interaction variable between large deposits and the FDIC insurance limit raise still dampens the hazard-reducing effect of the introduction of the FDIC raise for banks with a high percentage of large deposits. A

possible interpretation may be that depositors with accounts holding between USD 100,000 and USD 250,000 profited from the raise, but customers holding deposits in excess of USD 250,000 might have interpreted the intervention as a warning sign. Further, the more a bank was invested in residential real estate the lower its hazard, which may be counterintuitive given the financial crisis has its roots in the real estate sector. Lastly, it should be noted that the significant effect in assets is mainly driven by the failures of the three largest banks; excluding them from the analysis leads to statistically insignificant coefficients on assets (not shown in the table).

The appendix further lists results including censored Google series in Table C.3 as well as results ignoring weighting. The coefficients confirm the results shown above. The coefficient on the Google variable is attenuated towards zero when using censored Google series, which is expected as falls in the Google Index are overstated in the censored case.

Table 4.6 presents results for forecasting where the contemporaneous $Google_{it}$ growth variable is replaced with variables that are lagged by two to five weeks or with growth rates spanning two to five weeks. Again, the dataset containing only uncensored Google series is used. The control variables remain the same, but the output table is restricted to the coefficients for Google variables only. The top half presents specifications including lagged values of the Google variable from two up to five weeks, with the last column including all lags. The size of the Google coefficient goes toward zero and becomes statistically insignificant as one moves back in time. An exception is the last column in the upper half including all lags, showing significant effects for the three weeks before failure, confirming the pattern seen initially in Figure 4.2.

In the bottom half of Table 4.6, the Google variable covers the accumulated growth rate over a longer period, from a 2-week period up to a 5-week period. The results confirm the previous statements: Results are mainly driven by the Google growth values in the week of the failure; adding additional weeks and lengthening of the time period barely changes the estimated coefficients.

4.5 Conclusion

Washington Mutual was still considered well capitalized shortly before its closure (Grind, 2012), but the situation changed rapidly in mid-September. Within a few weeks, a well-capitalized bank - which admittedly did have problems with its mortgages - had to be shut down, as closing the bank was apparently the only option to stop the ongoing run on deposits. Once a bank run is kicked off, a vicious feedback-loop is started and withdrawals spread like a virus. As the Diamond and Dybvig (1983) model shows, one ends up in an equilibrium where it becomes rational for every agent to pull out their funds; even deposit insurance may prove ineffective at this point (Iyer and Puria, 2012; Grind, 2012).

Such bank failures are hard to predict. Empirical research analyzing the survival and survival time of banks by making use of their balance sheets provides insights, but these studies have their limits when it comes to the timing of the failure. Other research focussing on single banks helps understanding the dynamics during a bank run, but cannot explain when or why the bank run occurred in the first place. Google data can provide additional insights and accuracy in this field. As this study demonstrates, Google search volumes start rising up to two weeks before failure, indicating increased attention on the internet for an individual bank. By capturing short term dynamics that cannot be reflected in quarterly balance sheet and revenue data, Google queries can be a valuable improvement to more traditional predictions, especially when it comes to the timing of the failure. Compared to other instruments used to capture publicly available information, Google has the advantage of being "democratically" weighted rather than binary or influenced by other variables, reflecting the spread of information more accurately. While it is hard to know how many readers read a newspaper article, a rising Google index can always be translated into more people being concerned.

Nevertheless, it should be pointed out that Google search queries have their limits, too. First and foremost, one does not know what drives the spike in search queries or what

actions follow after the Google search, making any causal claims hard to defend. Whether a news article leads to the rising search volume or customers looking for their e-banking accounts is unknown. It would be rash to equate a rising Google Index with a bank run. What this study shows is that it can serve as a warning signal that failure is imminent. Still, the timing of spikes in search volumes remains hard to predict. As one moves further away from the bank's failure date by more than three weeks, Google loses its predictive power.

One should also keep in mind Google data's technical limitations. First, the data are censored. Particularly small banks fail to pass the search volume threshold, which means that there is no data available at all. Second, Google publishes only relative numbers, which allows for the use of growth rates only. Third, Google is not the internet. Google may be a popular search engine, but it does not track all activity on the web. Instead of a substitute, Google data should therefore be seen as a supplement to balance sheet and revenue analysis.

References

Askitas, N. and K. F. Zimmermann (2009). Google econometrics and unemployment forecasting. Technical report, German Institute for Economic Research.

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Breyer, B. N., S. Sen, D. S. Aaronson, M. Stoller, B. A. Erickson, and M. L. Eisenberg (2011). Use of Google Insights for Search to track seasonal and geographic kidney stone incidence in the United States. *Urology*.

Calomiris, C. W. and J. R. Mason (2003). Fundamentals, panics, and bank distress during the depression. *American Economic Review*, 1615–1647.

Choi, H. (2009). Predicting initial claims for unemployment benefits. *SSRN 1659307*.

Choi, H. and H. Varian (2012). Predicting the present with Google Trends. *Economic Record 88*(s1), 2–9.

Cole, R. A. and J. W. Gunther (1995). Separating the likelihood and timing of bank failure. *Journal of Banking & Finance 19*(6), 1073–1089.

D'Amuri, F. and J. Marcucci (2010). Google it! Forecasting the US unemployment rate with a Google job search index. *FEEM Working Paper No. 31.2010*.

Della Penna, N. and H. Huang (2009). Constructing consumer sentiment index for US using Google searches. *Working Paper*.

Diamond, D. W. and P. H. Dybvig (1983). Bank runs, deposit insurance, and liquidity. *The Journal of Political Economy*, 401–419.

Donaldson, R. G. (1992). Costly liquidation, interbank trade, bank runs and panics. *Journal of Financial Intermediation 2*(1), 59–82.

Federal Deposit Insurance Corp. Bank data and statistics. `http://www.fdic.gov/bank/statistical/`.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2008). Detecting influenza epidemics using search engine query data. *Nature 457*(7232), 1012–1014.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences 107*(41), 17486–17490.

Google. Google Insights for Search. `http://www.google.com/insights/search/`.

Gorton, G. (1988). Banking panics and business cycles. *Oxford economic papers 40*(4), 751–781.

Gorton, G. and A. Winton (2003). Financial intermediation. *Handbook of the Economics of Finance 1*, 431–552.

Grind, K. (2012). *The Lost Bank: The Story of Washington Mutual-The Biggest Bank Failure in American History*. Simon & Schuster.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement 36*(3), 119–167.

Iyer, R., M. Puri, and N. Ryan (2013). Do depositors monitor banks? Technical report, National Bureau of Economic Research.

Iyer, R. and M. Puria (2012). Understanding bank runs: the importance of depositor-bank relationships and networks. *The American Economic Review 102*(4), 1414–1445.

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*.

Kelly, M. and C. O Grada (2000). Market contagion: Evidence from the panics of 1854 and 1857. *American Economic Review*, 1110–1124.

Lancaster, T. (1992). *The econometric analysis of transition data*. Number 17. Cambridge University Press.

Los Angeles Times (2008, June 28). Senator asks regulators to probe the financial health of Indymac.

Mathiesen, J., L. Angheluta, P. T. H. Ahlgren, and M. H. Jensen (2013). Excitable human dynamics driven by extrinsic events in massive communities. *Proceedings of the National Academy of Sciences*.

McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of affective disorders 122*(3), 277–279.

McLaren, N. and R. Shanbhogue (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin No. 2011 Q2*.

Moat, H. S., C. Curme, A. Avakian, D. Y. Kenett, E. H. Stanley, and T. Preis (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports 3*.

O Grada, C. and E. N. White (2003). The panics of 1854 and 1857: A view from the emigrant industrial savings bank. *The Journal of Economic History 63*(1), 213–240.

Office of Thrift Supervision (2008, September 25). OTS fact sheet on Washington Mutual Bank.

Preis, T., H. S. Moat, and E. H. Stanley (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports 3*.

Preis, T., D. Reith, and H. Stanley (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368*(1933), 5707–5719.

Radinsky, K., S. Davidovich, and S. Markovitch (2008). Predicting the news of tomorrow using patterns in web search queries. In *Web Intelligence and Intelligent Agent Technology*, Volume 1, pp. 363–367. IEEE.

Rigobon, R. and T. M. Stoker (2007). Estimation with censored regressors: Basic issues. *International Economic Review 48*(4), 1441–1467.

Saunders, A. and B. Wilson (1996). Contagious bank runs: Evidence from the 1929–1933 period. *Journal of Financial Intermediation 5*(4), 409–423.

Seabrook, A. (2008, July 12). Interview with Burt Ely. *National Public Radio*.

Shin, H. S. (2009). Reflections on Northern Rock: the bank run that heralded the global financial crisis. *The Journal of Economic Perspectives*, 101–120.

Tefft, N. (2011). Insights on unemployment, unemployment insurance, and mental health. *Journal of Health Economics*.

United States Census Bureau. 2010 Census urban and rural classification and urban area criteria. `http://www.census.gov/geo/www/ua/2010urbanruralclass.html`.

Whalen, G. (1991). A proportional hazards model of bank failure: an examination of its usefulness as an early warning tool. *Federal Reserve Bank of Cleveland Economic Review 27*(1), 21–31.

Wheelock, D. and P. Wilson (2000). Why do banks disappear? The determinants of US bank failures and acquisitions. *Review of Economics and Statistics 82*(1), 127–138.

Wu, L. and E. Brynjolfsson (2013). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economics of Digitization*. University of Chicago Press.

Tables and Figures

Table 4.1: Overview of samples used

Sample	**Observations**	**Failing banks**	**Surviving banks**
Original sample	180,291	428	387
With Google series	45,835	115	95
With uncensored Google series	10,296	25	48

Table 4.2: Bank failures over time

	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
No. of failures	3	0	0	3	25	140	157	92	16	436

The year 2012 covers only the first quarter of the year.

The years 2004, 2005, 2006 are excluded from the analysis.

Source: FDIC

Table 4.3: Summary statistics

Full sample

		Failures				*Survivors*		
Variable	**Mean**	**St. Dev.**	**Min**	**Max**	**Mean**	**St. Dev.**	**Min**	**Max**
Google data dummy	.269	-	0	1	0.245	-	0	1
Gross assets	1'518.531*	15,727.860	0	322,059.800	394.069*	1,088.604	3.220	12,585.320
(in USD millions)								
Capital	7.502*	3.213	0	22.828	12.411*	6.076	0	77.759
Troubled assets	8.252*	5.293	0	42.943	3.364*	2.928	0.290	17.944
Net income	-0.498*	0.333	-2.729	0.312	0.130*	0.286	-1.073	2.317
Securities	8.191*	6.634	0	40.711	21.743*	14.452	0	76.510
Large CDs	16.209	8.164	0	53.467	16.173	7.440	0	39.799
Insider	1.093	1.389	0	10.956	1.345	1.488	0	10.716
Holding Co.	0.702	-	0	1	.692	-	0	1
Entering age	35.665*	38.406	0.071	156.493	68.375*	43.438	0.186	170.012
Urban	0.341*	-	0	1	0.437*	-	0	1
Observations	428				387			

With Google data

		Failures				*Survivors*		
Variable	**Mean**	**St. Dev.**	**Min**	**Max**	**Mean**	**St. Dev.**	**Min**	**Max**
Google growth rate	2.699*	2.331	-8.673	6.960	*1.632	2.515	-11.384	5.146
Google data dummy	1	0	1	1	1	0	1	1
Gross assets	4,618.891*	30,184.530	40.085	322,059.800	727.244*	1,803.306	20.207	12,585.320
(in USD millions)								
Capital	9.996*	2.502	5.522	22.828	11.976*	7.248	6.540	77.759
Troubled assets	10.265*	6.488	1.191	42.943	3.849*	3.417	0.420	17.944
Net income	-0.619*	0.426	-2.729	0.057	0.155*	0.344	-0.613	2.317
Securities	11.212*	7.978	0	40.711	20.825*	14.152	0	73.761
Large CDs	20.869*	7.916	5.282	53.467	17.598*	7.744	0	39.799
Insider	1.124	1.639	0	10.956	1.321	1.214	0	4.966
Holding Co.	0.687	-	0	1	0.726	-	0	1
Entering age	38.656*	38.511	0.624	156.493	60.966*	46.139	1.572	144.471
Urban	0.304	-	0	1	0.474	-	0	1
Observations	115				95			

Uncensored Google series only

		Failures				*Survivors*		
Variable	**Mean**	**St. Dev.**	**Min**	**Max**	**Mean**	**St. Dev.**	**Min**	**Max**
Google growth rate	1.313	1.396	-0.024	6.857	0.791	1.010	0.127	5.146
Google data dummy	1	0	1	1	1	0	1	1
Gross assets	15,400.920*	64,138.350	60.974	322,059.800	405.449*	574.211	36.744	2,316.473
(in USD millions)								
Capital	10.119	2.769	5.522	15.224	11.684	2.164	8.894	18.799
Troubled assets	9.868*	6.794	2.388	30.363	3.971*	2.852	0.878	11.329
Net income	-0.718*	0.512	-2.729	-0.152	0.109*	0.225	-0.467	0.526
Securities	10.194*	6.784	1.575	28.671	20.281*	12.676	1.964	47.787
Large CDs	22.849	10.780	5.701	53.467	15.693	5.719	7.828	27.899
Insider	1.514	2.707	0	10.956	1.418	1.134	0	3.969
Holding Co.	.560	-	0	1	.696	-	0	1
Entering age	28.636*	27.770	0.953	99.806	69.443*	47.741	3.773	143.411
Urban	0.360	-	0	1	0.609	-	0	1
Observations	25				23			

Sources: Google Insights for Search, FDIC, Bloomberg, 2010 US Census.
Reported values are averaged by institution and cover the period from Jan 2007 to Mar 2012 or up to failure, respectively.
An $*$ indicates that the Wilcoxon rank-sum test statistic for a shift in the location parameter between the two groups is significant at the one percent level.
The difference in gross assets in the lower third of the table (uncensored Google series only) is not signifcant anymore if the two largest banks Washington Mutual and Indymac are excluded.
ComRE, ResRE and Insdep variables have been omitted.

Table 4.4: Results on survival and Google data availability

	Google only	**PWC-FDIC**	**PWC-yearly**	**PWC-quarterly**	**Cox PH**
Variable	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
Google	0.120	0.315	0.289	0.273	0.234
	(0.157)	(0.232)	(0.234)	(0.232)	(0.236)
Capital	-	-0.470***	-0.475***	-0.474***	-0.489***
		(0.055)	(0.028)	(0.029)	(0.030)
Troubledassets	-	0.055***	0.055***	0.057***	0.058***
		(0.007)	(0.007)	(0.007)	(0.007)
Netincome	-	0.067	0.054	0.049	0.033
		(0.047)	(0.043)	(0.041)	(0.038)
Securities	-	-0.064***	-0.061***	-0.060***	-0.058***
		(0.014)	(0.014)	(0.014)	(0.014)
LargeCDs	-	-0.094**	-0.093**	-0.101**	-0.091*
		(0.036)	(0.030)	(0.038)	(0.041)
Insdep	-	0.090*	0.088**	0.097*	0.089*
		(0.037)	(0.030)	(0.039)	(0.041)
ComRE	-	0.034**	0.035**	0.035**	0.036**
		(0.011)	(0.011)	(0.011)	(0.0011)
ResRE	-	-0.000	0.001	-0.000	-0.001
		(0.010)	(0.010)	(0.010)	(0.010)
Insider	-	-0.047	-0.054	-0.055	-0.075
		(0.060)	(0.060)	(0.060)	(0.074)
Assets	-	0.078	0.076	0.074	0.072
		(0.074)	(0.074)	(0.074)	(0.074)
Age	-	-0.001	-0.001	-0.001	-0.002
		(0.003)	(0.003)	(0.003)	(0.003)
Holding	-	-0.705**	-0.703**	-0.690**	-0.664**
		(0.237)	(0.241)	(0.243)	(0.248)
Urban	-	0.436*	0.419*	0.427*	0.444*
		(0.191)	(0.192)	(0.190)	(0.191)
LOIS	-	0.180	0.054	-0.360	-
		(0.497)	(0.194)	(0.380)	
Piecewise constant haz.	quarterly	2-period	yearly	quarterly	-
Google Variable	Dummy	Dummy	Dummy	Dummy	Dummy
Observations	181,113	181,113	181,113	181,113	181,113
Subjects	818	818	818	818	818
Failures	428	428	428	428	428
Weighted failures	40.330	40.330	40.330	40.330	40.330
Log-pseudolikelihood	-169.566	-33.191	-32.762	-31.821	-104.610

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

Clustered standard errors reported in parentheses (clustered on subject).
Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.
Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.
Episodes are split on a weekly basis.
The Cox PH model uses the Breslow method for ties.

Table 4.5: Main results, uncensored Google series only

	Google only	**PWC-FDIC**	**PWC-yearly**	**PWC-quarterly**	**Cox PH**
Variable	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
Google	0.042***	0.024*	0.024***	0.042*	0.048***
	(0.012)	(0.010)	(0.007)	(0.016)	(0.010)
Capital	-	-0.506***	-0.539***	-0.579***	-0.625***
		(0.094)	(0.090)	(0.107)	(0.160)
Troubledassets	-	0.049†	0.059*	0.070**	0.077**
		(0.028)	(0.028)	(0.024)	(0.026)
Netincome	-	0.083	0.072	-0.076	-0.050
		(0.066)	(0.069)	(0.130)	(0.130)
Securities	-	-0.248**	-0.261**	-0.317**	-0.348**
		(0.086)	(0.092)	(0.119)	(0.134)
LargeCDs	-	-0.342	-0.237	-1.453**	-5.083***
		(0.222)	(0.217)	(0.517)	(1.212)
Insdep	-	0.413†	0.304	1.515**	5.155***
		(0.231)	(0.219)	(0.502)	(1.204)
ComRE	-	-0.046	-0.055	-0.026	-0.027
		(0.044)	(0.051)	(0.053)	(0.051)
ResRE	-	-0.068***	-0.073***	-0.094*	-0.113*
		(0.016)	(0.022)	(0.037)	(0.047)
Insider	-	0.105	0.166	0.176	0.153
		(0.140)	(0.153)	(0.133)	(0.138)
Assets	-	1.021*	1.102*	0.826***	0.737**
		(0.436)	(0.508)	(0.240)	(0.268)
Age	-	-0.025	-0.029†	-0.030*	-0.023†
		(0.016)	(0.017)	(0.014)	(0.012)
Holding	-	2.007*	2.147*	1.056	0.635
		(0.940)	(1.019)	(0.787)	(0.935)
Urban	-	0.881	0.996	2.823**	2.976**
		(0.879)	(1.018)	(1.002)	(1.028)
LOIS	-	0.571*	0.322	0.126	-
		(0.269)	(0.807)	(1.102)	
Piecewise constant haz.	quarterly	2-period	yearly	quarterly	-
Google Variable	%-change	%-change	%-change	%-change	%-change
Observations	10,296	10,296	10,296	10,296	10,296
Subjects	48	48	48	48	48
Failures	25	25	25	25	25
Weighted failures	2.367	2.367	2.367	2.367	2.367
Log-pseudolikelihood	-8.249	1.517	1.653	2.863	2.366

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

Clustered standard errors reported in parentheses (clustered on subject).
Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.
Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.
Episodes are split on a weekly basis.
The Cox PH model uses the Breslow method for ties.

Table 4.6: Forecasting

Variable	**PWC**	**PWC**	**PWC**	**PWC**	**PWC**
Variable	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
1-week lag	-	-	-	-	0.039***
					(0.011)
2-week lag	.027	-	-	-	0.036**
	(0.019)				(0.012)
3-week lag	-	0.013	-	-	0.038*
		(0.014)			(0.019)
4-week lag	-	-	-.006	-	0.023
			(0.017)		(0.021)
5-week lag	-	-	-	0.007	0.018
				(0.011)	(0.013)
Piecewise constant haz.	quarterly	quarterly	quarterly	quarterly	quarterly
Google Variable	%-change	%-change	%-change	%-change	%-change
Observations	10,296	10,296	10,296	10,296	10,296
Subjects	48	48	48	48	48
Failures	25	25	25	25	25
Weighted failures	2.367	2.367	2.367	2.367	2.367
Log-pseudolikelihood	2.353	2.219	2.191	2.195	3.243
2-week period	0.025***	-	-	-	
	(0.004)				
3-week period	-	0.021***	-	-	
		(0.003)			
4-week period	-	-	0.022***	-	
			(0.003)		
5-week period	-	-	-	0.024***	
				(0.003)	
Piecewise constant haz.	quarterly	quarterly	quarterly	quarterly	
Google Variable	%-change	%-change	%-change	%-change	
Observations	10,296	10,296	10,296	10,296	
Subjects	48	48	48	48	
Failures	25	25	25	25	
Weighted failures	2.367	2.367	2.367	2.367	
Log-pseudolikelihood	3.067	3.185	3.105	3.181	

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

Clustered standard errors reported in parentheses (clustered on subject).
Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.
Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.
Episodes are split on a weekly basis.

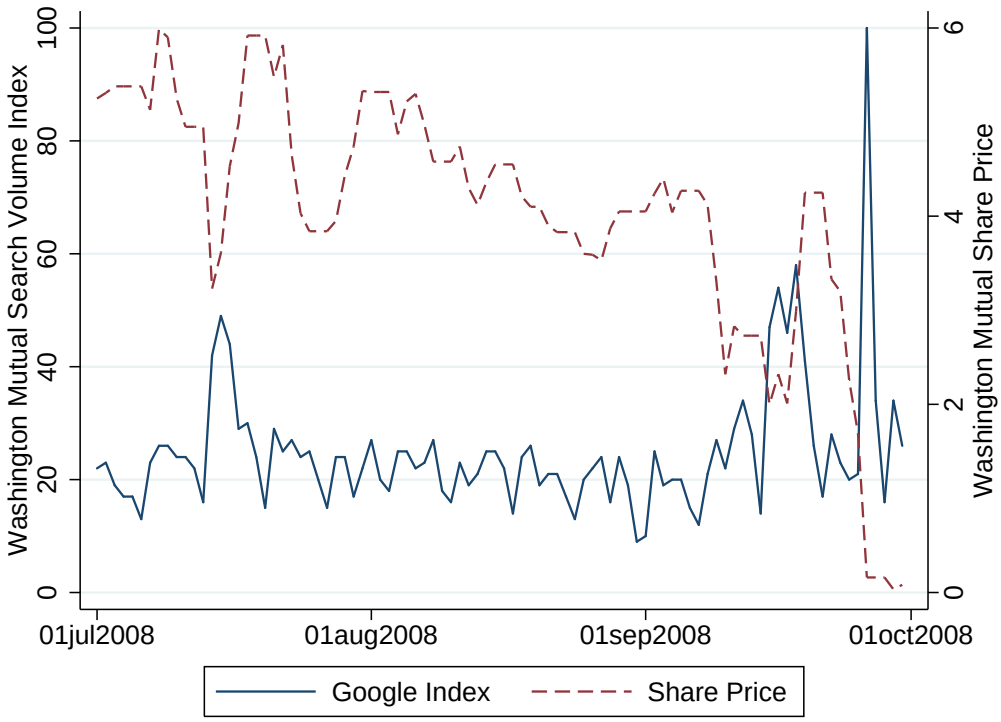


Figure 4.1: Google Search Volume Index and share price for “Washington Mutual“

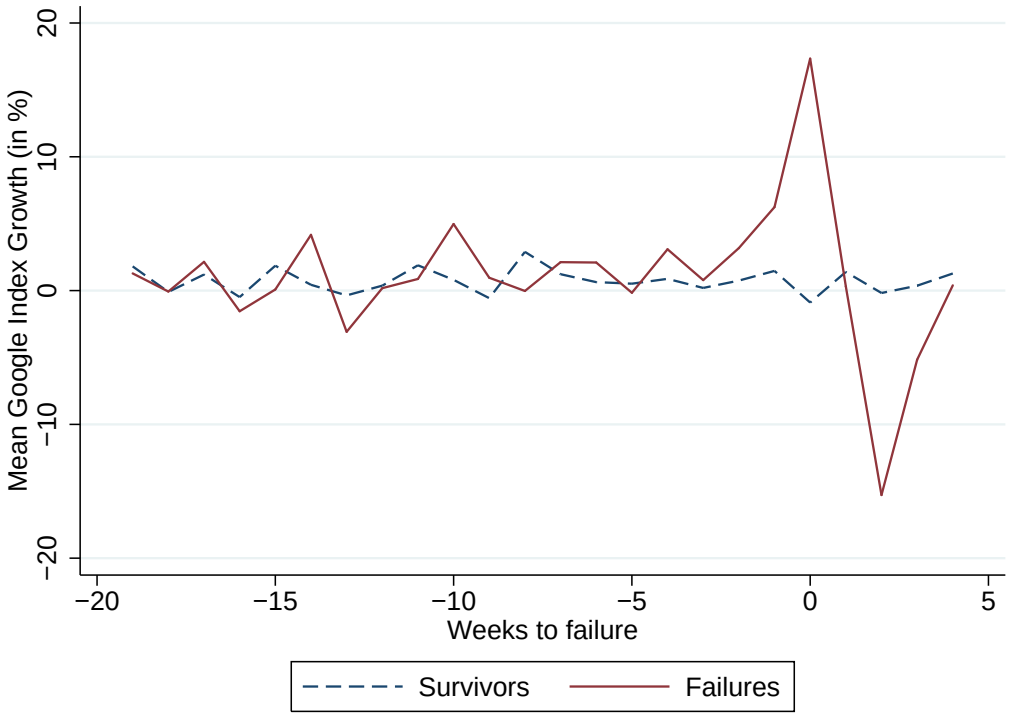


Figure 4.2: Google growth rates in the weeks prior to failure

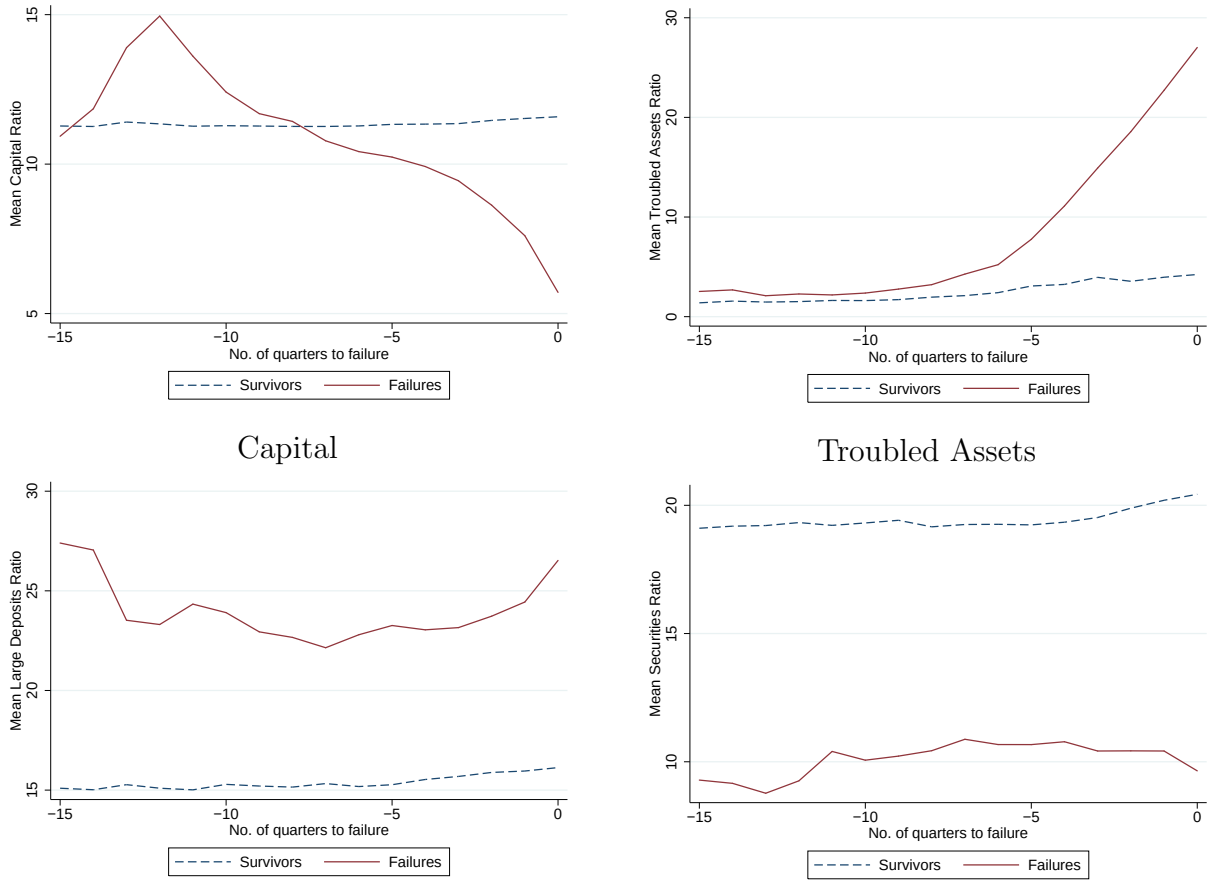


Figure 4.3:
Key balance sheet positions before failure, conditioned on observing uncensored Google series

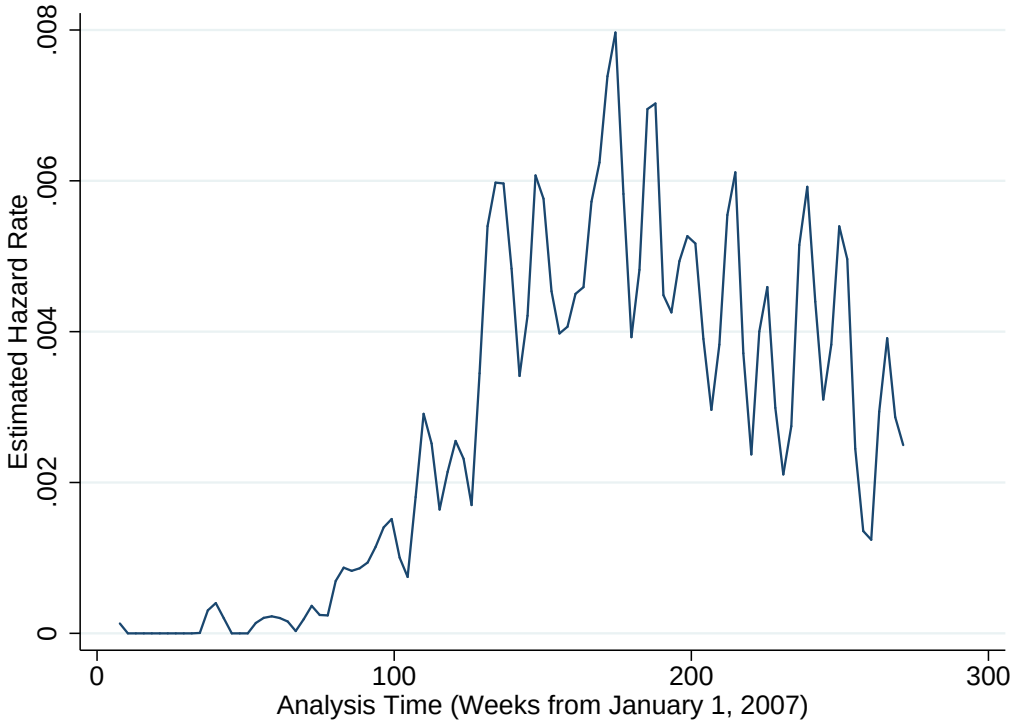


Figure 4.4: Smoothed hazard rate estimate

Appendix A

Chapter 2

Table A.1: Summary statistics on binary HI/BYE decisions, by gender

	Mean	Std. Dev.	Min	Max
Males				
Overall	0.498	0.500	0.000	1.000
Between		0.309	0.000	1.000
Within		0.401	-0.492	1.488
Observations	821,525			
Individuals	10,497			
$\bar{r}$	78.263			
Females				
Overall	0.141	0.348	0.000	1.000
Between		0.153	0.000	1.000
Within		0.323	-0.848	1.131
Observations	453,575			
Individuals	5,735			
$\bar{r}$	79.089			

Source: BLINQ; own calculations. Based on the 100 first decisions of all users.

Table A.2: Linear probability model results on preference estimates

	Female		Male	
	Coeff.	*SE*	*Coeff.*	*SE*
DV: Binary willingness-to-date decision				
Attractiveness, *standardized*	0.087	0.001	0.193	0.001
Squared diff. in attractiveness, *positive*	0.017	0.001	0.001	0.000
Squared diff. in attractiveness, *negative*	-0.002	0.000	-0.020	0.001
Years $\geq$ 18, *absolute*	0.005	0.000	-0.006	0.000
Years $<$ 18, *absolute*	-0.046	0.004	0.026	0.003
Squared diff. in age, *positive*	-0.001	0.000	-0.001	0.000
Squared diff. in age, *negative*	-0.001	0.000	-0.000	0.000
University	0.002	0.002	-0.003	0.002
Both university	0.006	0.006	0.010	0.005
Same school	0.002	0.002	0.005	0.002
German speaking candidate	-0.007	0.003	-0.002	0.002
Both German speaking	0.008	0.003	0.005	0.003
No. of friends, *in hundreds*	-0.000	0.000	0.001	0.000
Mutual friends	0.004	0.000	0.001	0.000
Squared diff. in friends, *positive*	-0.001	0.000	-0.000	0.000
Squared diff. in friends, *negative*	-0.000	0.000	-0.000	0.000
Distance in *km*	-0.000	0.000	-0.000	0.000
TRX	-0.000	0.000	0.000	0.000
Observations	453,575		821,525	
Individuals	5,735		10,497	
R^2 - overall	0.054		0.079	
R^2 - within	0.084		0.132	
R^2 - between	0.003		0.008	

Source: BLINQ; own calculations. Based on the 100 first decisions of all users.
Variable definitions as before.
Variables other than direct user-candidate comparisons relate to the candidate, not the user. User characteristics are captured in the fixed effect.

Table A.3: Fixed effects logit results on preference estimates (marginal effects)

	Female		Male	
	Coeff.	*SE*	*Coeff.*	*SE*
DV: Binary willingness-to-date decision				
Attractiveness, *standardized*	0.213	0.003	0.327	0.002
Squared diff. in attractiveness, *positive*	-0.001	0.004	-0.010	0.001
Squared diff. in attractiveness, *negative*	-0.016	0.000	-0.029	0.001
Years ≥ 18, *absolute*	0.012	0.001	-0.010	0.000
Years < 18, *absolute*	-0.146	0.010	0.043	0.006
Squared diff. in age, *positive*	-0.003	0.000	-0.001	0.000
Squared diff. in age, *negative*	-0.003	0.000	-0.000	0.000
University	0.009	0.004	-0.004	0.003
Both university	0.012	0.013	0.017	0.010
Same school	0.005	0.005	0.008	0.003
German speaking candidate	-0.021	-0.007	-0.004	0.004
Both German speaking	0.020	0.007	0.010	0.005
No. of friends, *in hundreds*	0.000	0.000	0.001	0.000
Mutual friends	0.007	0.001	0.001	0.000
Squared diff. in friends, *positive*	-0.001	0.001	-0.001	0.000
Squared diff. in friends, *negative*	-0.000	0.000	-0.001	0.000
Distance in *km*	-0.000	-0.000	-0.000	0.000
TRX	-0.001	0.000	0.000	0.000

Source: BLINQ; own calculations. Based on the 100 first decisions of all users. For females, 368 users (11,785 observations) were dropped because of all positive or all negative outcomes. For males, 947 users (35,589 observations) were dropped because of all positive or all negative outcomes. Marginal effects are calculated assuming a fixed effect of zero. Females: $Pr(y = 1|\text{FE is zero}) = 0.645$; males: $Pr(y = 1|\text{FE is zero}) = 0.463$. Discrete effects calculated for dummy variables.

Attractiveness is defined as the ratio of the number of *HI*'s a user got, divided by the number of times the user has been rated. The measure is standardized within gender. Differences are taken over the standardized measure. *Acceptancerate* is defined as the ratio between the number of times a user rates *HI*, divided by the total number of decisions she has taken. Standardization as in the case of attractiveness. *Age* is measured in years and bounded between 13 and 40 and is reformulated as the absolute difference from 18. *University* is a dummy indicating whether the user has a university listed on his Facebook profile. *Both university* is a dummy indicating whether *university* == 1 for both the user as well as the candidate.*Same school* is a dummy indicating whether both user and candidate list the same school on their Facebook profile. *German speaking* is a dummy indicating the language set in the app. *No. of friends* is the number of Facebook friends measured in hundreds, *mutual friends* the number of mutual friends that also use the dating application. *Squared difference in friends* is measured in units of 100,000. *Distance* is the distance in *km* between user and candidate, where the information on location was drawn just once, assuming users do not move. Only candidates within a 300*km* radius are considered. *Duration* is the fraction of the current decision number divided by the total number of decisions taken by a user.

Variables other than direct user-candidate comparisons relate to the candidate, not the user.

Table A.4: Fixed effects logit results on preference estimates (robustness I)

	Female		Male	
	Coeff.	*SE*	*Coeff.*	*SE*
DV: Binary willingness-to-date decision				
Attractiveness, *standardized*	1.045	0.011	1.251	0.008
Squared diff. in attractiveness, *positive*	0.040	0.016	-0.039	0.003
Squared diff. in attractiveness, *negative*	-0.086	0.002	-0.114	0.004
Years ≥ 18, *absolute*	0.048	0.004	-0.036	0.002
Years < 18, *absolute*	-0.625	0.046	0.060	0.025
Squared diff. in age, *positive*	-0.013	0.001	-0.005	0.000
Squared diff. in age, *negative*	-0.013	0.001	0.001	0.000
University	0.026	0.019	-0.006	0.013
Both university	0.168	0.069	0.130	0.042
Same school	0.013	0.024	0.023	0.014
German speaking candidate	-0.161	0.033	-0.023	0.017
Both German speaking	0.174	0.036	0.060	0.020
No. of friends, *in hundreds*	0.002	0.002	0.005	0.001
Mutual friends	0.044	0.004	0.004	0.002
Squared diff. in friends, *positive*	-0.006	0.003	-0.005	0.002
Squared diff. in friends, *negative*	-0.002	0.001	-0.002	0.001
Distance in *km*	-0.000	0.000	0.000	0.000
TRX	-0.000	0.000	0.005	0.000
Observations	411,214		730,067	
Individuals	5,196		9,585	
Log-likelihood	-97,327		-289,379	

Source: BLINQ; own calculations. Based on the 100 randomly drawn decisions of all users. For females, 517 users (22,479 observations) were dropped because of all positive or all negative outcomes. For males, 958 users (26,841 observations) were dropped because of all positive or all negative outcomes.

Variables are defined as previously.

Variables other than direct user-candidate comparisons relate to the candidate, not the user. User characteristics are captured in the fixed effect.

Table A.5: Fixed effects logit results on preference estimates (robustness II)

	Female		Male	
	Coeff.	*SE*	*Coeff.*	*SE*
DV: Binary willingness-to-date decision				
Attractiveness, *standardized*	0.922	0.060	1.318	0.042
Squared diff. in attractiveness, *positive*	-0.038	0.085	-0.034	0.017
Squared diff. in attractiveness, *negative*	-0.063	0.011	-0.092	0.023
Years $\geq$ 18, *absolute*	0.009	0.024	-0.032	0.010
Years $<$ 18, *absolute*	-0.640	0.227	0.566	0.112
Squared diff. in age, *positive*	-0.023	0.006	-0.002	0.001
Squared diff. in age, *negative*	-0.008	0.003	0.001	0.001
University	-0.097	0.111	0.046	0.073
Both university	-0.085	0.375	0.466	0.252
Same school	0.062	0.121	0.061	0.077
German speaking candidate	-0.121	0.182	-0.083	0.098
Both German speaking	0.252	0.200	0.157	0.113
No. of friends, *in hundreds*	-0.001	0.012	-0.012	0.008
Mutual friends	0.028	0.018	0.018	0.014
Squared diff. in friends, *positive*	0.010	0.028	-0.026	0.009
Squared diff. in friends, *negative*	-0.003	0.004	0.003	0.003
Distance in *km*	0.000	0.001	0.001	0.000
TRX	-0.001	0.001	-0.002	0.001
Observations	13,764		26,031	
Individuals	155		294	
Log-likelihood	-3,138		-9,896	

Source: BLINQ; own calculations. Based on the full decision history of 175 randomly drawn users. For females, 20 users (1,335 observations) were dropped because of all positive or all negative outcomes. For males, 37 users (1,891 observations) were dropped because of all positive or all negative outcomes.
Variables are defined as previously.
Variables other than direct user-candidate comparisons relate to the candidate, not the user. User characteristics are captured in the fixed effect.

Table A.6: Robustness results on best rank: all users

	Female		*Male*		*Female*		*Male*	
	Coeff.	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*
DV: $\ln bestrank$								
$\ln N$	0.061	(0.014)	0.547	(0.010)	-0.014	(0.014)	0.658	(0.008)
$\ln attract$					-1.403	(0.051)	-1.360	(0.014)
$\ln accrate$					-0.513	(0.023)	-0.056	(0.014)
Constant	1.591	(0.097)	0.683	(0.076)	-0.359	(0.109)	-4.138	(0.085)
Observations	5,114		8,232		5,114		8,232	
R^2	0.003		0.218		0.177		0.706	
F-Stat	18.35		2,721		344.5		4,854	

Source: BLINQ; own calculations.

The sample considers the best-ranked matched mate of all users, including still actively who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match. The dependent variable *lnrank* is the logarithm of the individual-specific rank of the matched mate, where the rank is based on the estimated preference parameters reported previously. $\ln N$ is the logarithm of the length of an individual's search sequence, i.e., the number of decisions a user has taken. $\ln attract$ and $\ln select$ are the logarithmized measures of *attractiveness* and *acceptancerate* reported previously.

Table A.7: Robustness results on best rank: $N \geq 100$

	Female		*Male*		*Female*		*Male*	
	Coeff.	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*
DV: $\ln bestrank$								
$\ln N$	-0.031	(0.030)	0.529	(0.021)	-0.102	(0.028)	0.622	(0.015)
$\ln attract$					-1.521	(0.076)	-1.202	(0.022)
$\ln accrate$					-0.557	(0.034)	-0.028	(0.022)
Constant	2.330	(0.198)	0.872	(0.147)	0.025	(0.209)	-3.416	(0.150)
Observations	2,416		3,082		2,416		3,082	
R^2	0.001		0.156		0.183		0.627	
F-Stat	1.13		606.93		188.4		1,331	

Source: BLINQ; own calculations.

The sample considers the best-ranked matched mate for users who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match and at search length $N \geq 100$. The dependent variable *lnrank* is the logarithm of the individual-specific rank of the matched mate, where the rank is based on the estimated preference parameters reported previously. $\ln N$ is the logarithm of the length of an individual's search sequence, i.e., the number of decisions a user has taken. $\ln attract$ and $\ln select$ are the logarithmized measures of *attractiveness* and *acceptancerate* reported previously.

Table A.8: Results on median rank (all observations)

	Female		*Male*	
	Coeff.	*SE*	*Coeff.*	*SE*
DV: $\ln medianrank$				
$\ln N$	1.081	(0.013)	0.986	(0.008)
$\ln attract$	-0.132	(0.041)	-0.313	(0.012)
$\ln accrate$	0.258	(0.020)	0.286	(0.015)
Constant	-1.302	(0.092)	-1.512	(0.073)
Observations	2,652		3,381	
R^2	0.781		0.876	
F-Stat	2,759		5,706	

Source: BLINQ; own calculations.

The sample considers the best-ranked matched mate for users who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match. The dependent variable *lnrank* is the logarithm of the individual-specific rank of the matched mate, where the rank is based on the estimated preference parameters reported previously. $\ln N$ is the logarithm of the length of an individual's search sequence, i.e., the number of decisions a user has taken. $\ln attract$ and $\ln select$ are the logarithmized measures of *attractiveness* and *acceptancerate* reported previously.

Table A.9: Search length descriptives

	Female		Male	
	Coeff	*St. Dev.*	*Coeff*	*St. Dev.*
ln *attract*	0.260	(0.052)	0.362	(0.019)
ln *accrate*	-0.489	(0.018)	-0.108	(0.020)
Age	0.041	(0.004)	0.059	(0.003)
Uni	0.061	(0.068)	0.100	(0.054)
Constant	4.517	(0.111) 6.086	(0.108)	
Observations	6,066		11,302	
R^2	0.138		0.074	

Source: BLINQ; own calculations.

Table A.10: First impressions in later stages, males

	Measures based on search length			Measures based on matches		
	Convstart	*Reply*	*Phone*	*Convstart*	*Reply*	*Phone*
Specification 1						
xb1	0.531	-0.003	0.131			
	(0.011)	(0.017)	(0.030)			
xb2	-0.013	-0.002	0.010			
	(0.003)	(0.005)	(0.086)			
sentmess			0.061			
			(0.005)			
recmess			0.052			
			(0.006)			
logL	-21,426	-7,625	-3,119			
Observations	49,979	27,354	28,699			
Specification 2						
rank1	-0.089	-0.002	-0.021	-0.013	-0.001	-0.003
	(0.002)	(0.003)	(0.005)	(0.000)	(0.001)	(0.001)
rank2	0.007	-0.005	-0.005	0.001	-0.004	-0.001
	(0.002)	(0.003)	(0.004)	(0.000)	(0.001)	(0.001)
rankdiffsq	-0.000	0.000	0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sentmess			0.062			
			(0.005)			
recmess			0.051			
			(0.006)			
logL	-21,582	-7,623	-3,119	-22,111	-7,573	-3,119
Observations	49,979	27,354	49,979	51,715	27,354	28,699
Specification 3						
pctile1	-2.613	0.086	-0.665	-1.875	0.025	-0.505
	(0.055)	(0.090)	(0.154)	(0.039)	(0.068)	(0.112)
pctile2	0.000	-1.025	-0.249	0.103	-0.914	-0.103
	(0.052)	(0.100)	(0.154)	(0.047)	(0.089)	(0.137)
sentmess			0.062			0.062
			(0.005)			(0.005)
recmess			0.051			0.052
			(0.006)			(0.006)
logL	-21,475	-7,570	-3,119	-21,500	-7,570	-3,119
Observations	49,979	27,354	28,699	49,979	27,354	28,699

Source: BLINQ; own calculations. Standard errors in parentheses.

The sample considers users who have been inactive for at least 90 days, restricting the sample to users who have finished their mate search. The sample includes all users with a match. *Convstart* is a dummy variable indicating whether a user starts a conversation, *reply* an indicator whether a user replies to a started conversation (conditional on the conversation being started). *Phone* is a dummy variable indicating whether a phone number was exchanged. $xb1$ and $xb2$ are the indices calculated according to estimated preference parameters for user and candidate, respectively. $rank1$ and $rank2$ are the ranks calculated based on the indices (in hundreds for the left half of the table), with rank1 equal to 1 being the most attractive candidate presented to the user. In the $rankdiffsq$ is the squared difference in ranks, measured in 10,000 units in the case of the left half of the table. $pct1$ and $pct2$ are the respective percentile ranks. *sentmess* is the number of messages sent to the candidate, *recmess* the number of messages received.

Observations number differ because of no variation at the individual level as well as bisexual candidates.

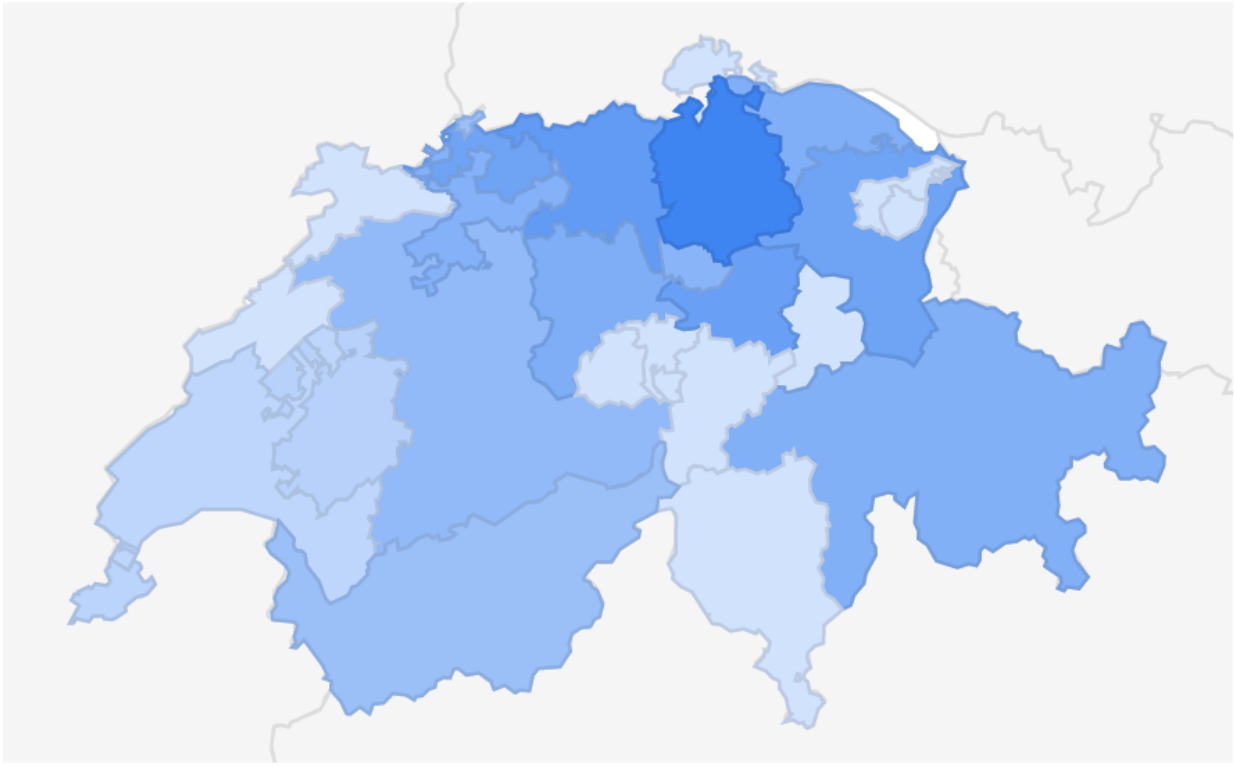
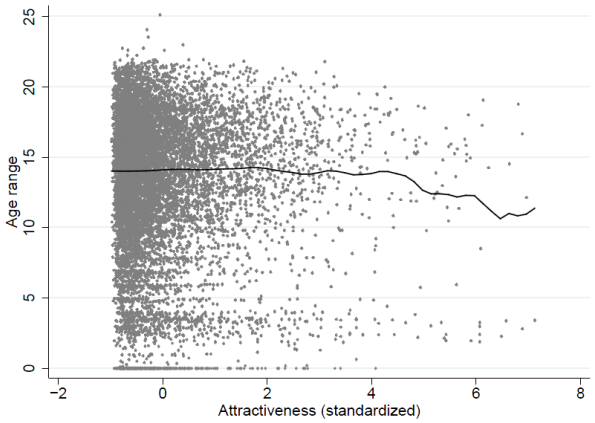
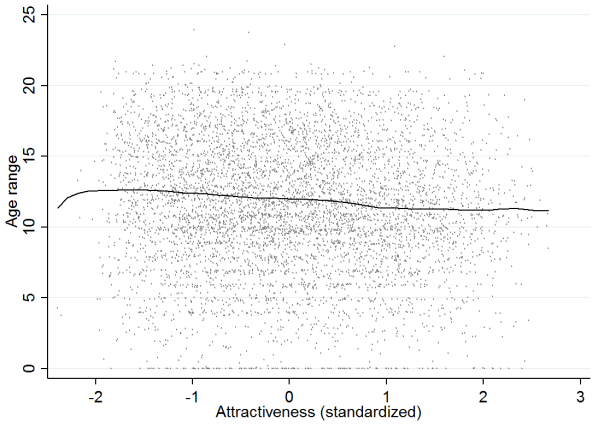


Figure A.1: Regional interest for BLINQ as measured by Google Trends data (1/2013 - 7/2015)

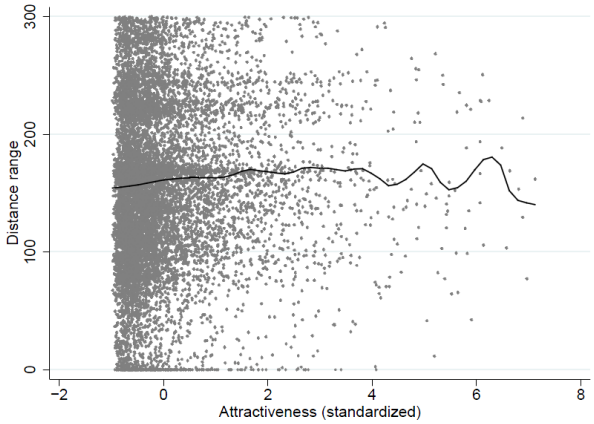


(a) Candidate age range (male)

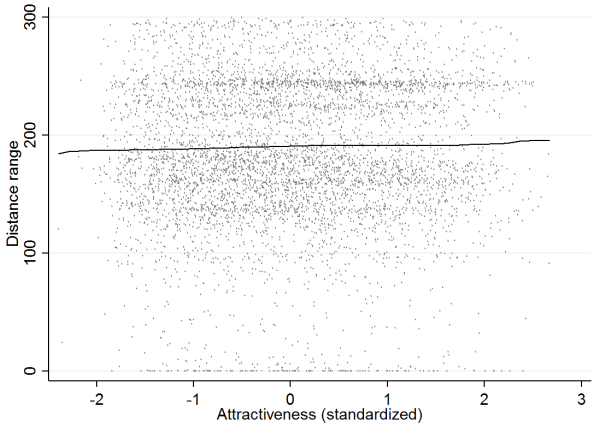


(b) Candidate age range (female)

Figure A.2: Age range of candidates in years, by gender (local polynomial fit)



(a) Candidate distance range (male)



(b) Candidate distance range (female)

Figure A.3: Distance range of candidates in km, by gender (local polynomial fit)

Appendix B

Chapter 3

Table B.1: Overdispersion in restaurant checkins at different levels of aggregation

Level of aggregation	*Global*	*ZIP*	*Price category*	*ZIP x Price*
Mean	282.42	282.42	282.42	282.42
SD	480.47	178.25	119.53	228.72
Skew	1.90	1.44	0.07	1.78
Kurtosis	42.87	5.35	1.05	6.20

Source : Yelp. 2015 data used, covering cumulative checkins across a 5 year period.

Appendix C

Chapter 4

Table C.1: Description of Variables

Variable	**Variable Name**	**Definition**
Googledata	Google Dummy	Dummy variable indicating whether Google data is available
Google growth	Google	Google growth rate (weekly; rate covering 1 to 5 weeks)
Capital	Capital	Ratio of equity capital and loan loss reserves to gross assets.
Troubled Assets	Troubledassets	Ratio of loans past due 90 days or more, nonaccrual loans, and other real estate owned to gross assets
Net Income	Netincome2	Ratio of net income to gross assets
Securities	Sec	Ratio of investment securities to gross assets
Large CDs	Largecds	Ratio of time deposits of USD 100'000 or more to gross assets
C&I Loans	Comindloans	Ratio of commercial and industrial loan to gross assets
Agricultural Loans	Agrloans	Ratio of agricultural production loans to gross assets
Commercial Real Estate Loans	Comrelos	Ratio of construction loans and loans secured by multifamily, nonresidential, or farm real estate to gross assets
Residential Real Estate Loans	Resire	Ratio of loans secured by 1-4 family real estate to gross assets
Consumer Loans	Consumer	Ratio of consumer loans to gross assets
Other Loans	Otherloans	Ratio of all other loans to gross assets
Insider Loans	Insider	Ratio of insider loans to gross assets
Salary Expenses	Salaries2	Ratio of salaries to equity capital
Premises Expense	Premise2	Ratio of expenses of premises to equity capital
Other Noninterest Expenses	Otnonint	Ratio of other nonintereste expenses to gross assets
Assets	Assetsize	Logarithm of gross assets (USD thousands)
Entering age	Enteringage	Age of the institution (years) when first entering the dataset
Holding Company	Hc_Dummy	Dummy variable to indicate whether the institution belongs to a holding company
Urban	Urban	One for urban counties, zero otherwise
Insurance	Insurance	Dummy variable to indicate the raise of the FDIC insurance limit in October 2008
Insdep	Insdep	Interaction of insurance dummy and large deposits ratio
LIBOR	LIBOR	3 month London Interbank Offered Rate
OIS	OIS	3 month Overnight Indexed Swap (OIS)
LOIS	LOIS	Difference between LIBOR and OIS as a measure of health of the banking system

No. of failures	Failures	Number of bank failures in a given week, to control for contagion effects.
Year	2008-2012	Dummies indicating year
Censoring dummy	Cens2	Dummy indicating temporarily censored Google series

Sources: Google Insights for Search, FDIC, Bloomberg, 2010 US Census.

Table C.2: Google Query Index Value Growth Rates

	Google Growth Rate				
Statistic	**1 week**	**2 weeks**	**3 weeks**	**4 weeks**	**5 weeks**
Observations	16323	16279	16235	16191	16147
Mean	1.349	1.523	1.325	1.772	1.947
Standard Deviation	16.696	17.750	18.702	18.552	19.095
Minimum	-66.250	-68.966	-64.286	-67.308	-69.091
Maximum	200.000	316.667	440.000	354.545	350.000

Includes only uncensored observations.

Source: Google Insights for Search

Table C.3: Additional results (censored Google series)

	(15)	(16)	(17)
Variable	**Coefficient**	**Coefficient**	**Coefficient**
Google	-	.025***	.017***
		(.007)	(.001)
Capital	-.500***	-.508***	-.674***
	(.030)	(.127)	(.058)
Troubledassets	.054***	.078†	.026*
	(.008)	(.045)	(.010)
Sec	-.059***	-.191†	-.076**
	(.014)	(.114)	(.024)
Largecds	-.122**	-.273†	-.160
	(.045)	(.153)	(.140)
Insurance	-1.460†	-2.008	-1.370
	(.773	(2.597)	(2.139)
Insdep	.107*	.324*	0.118
	(.046)	(.130)	(.141)
Comrelos	.042***	-.142**	.035
	(.012)	(.050)	(.024)
Resire	.003	-.154***	-.036†
	(.011)	(.031)	(.020)
Assets	.068	.929*	.101
	(.105)	(.427)	(.152)
Time Variables	Yes	No	Yes
Google Variable	Growth	Growth	Growth
Observations	280,736	16,323	54,545
Subjects	744.00	44.00	193.00
Failures	30.389	1.80	7.88
Log-pseudolikelihood	-0.000	3.393	7.167

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

Clustered standard errors reported in parentheses (clustered on subject).

Additional controls: netincome2, comindloans, agrloans, consumer, otherloans, insider, salaries2, premise2, otnonint, enteringage, hc_Dummy, urban, lois, failures, annual time dummies for the years 2008 to 2012.

Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.

Episodes are split on a weekly basis.

Curriculum Vitae

Florian Basil Schaffner

Born March 5^{th}, 1986
from Basel, BS
Swiss national

Education

2012 - 2017	Doctoral program at the *Zurich Graduate School of Economics*, University of Zurich, Switzerland
2011 - 2012	Exchange Semester at Tsinghua University, China
2009 - 2012	Master of Arts in *Economics*, University of Zurich, Switzerland
2006 - 2009	Bachelor of Arts in *International Relations*, University of Geneva, Switzerland

Professional experience

2012 - 2017	Research and Teaching assistant at the Department of Economics, University of Zurich, Switzerland
2012 - 2016	Contributor, PUNKT Magazin, Switzerland
2010 - 2014	Investment Analyst, Zurich Insurance, Switzerland
2009 - 2010	Contributor and Journalism Student, Ringier, Switzerland
2007 - 2008	Associate Operations, VZ Vermögenszentrum, Switzerland

July, 2017